DOCUMENT RESUME

ED 208 003 TM 810 675

AUTHOR	Glass, Gene V.; And Others
TITLE	Integration of Research Studies: Meta-Analysis of Research. Methods of Integrative Analysis; Final Report.
INSTITUTION	Colorado Univ., Boulder. Lab. of Educational Research.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	15 Aug 80
GRANT	NIE-G-78-0148
NOTE	340p.; Appendix B is removed due to copyright restrictions.
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	*Data Analysis; *Literature Reviews; *Research Methodology; Research Problems; Statistical Analysis
IDENTIFIERS	*Meta Analysis

ABSTRACT

Integrative analysis, or what is coming to be known as meta-analysis, is the integration of the findings of many empirical research studies of a topic. Meta-analysis differs from traditional narrative forms of research reviewing in that it is more quantitative and statistical. Thus, the methods of meta-analysis are merely statistical methods, suitably adapted in many instances, that are applicable to the job of integrating findings from many studies. A meta-analysis involves about a half-dozen steps: (1) defining the problem, (2) finding the research studies, (3) coding the study characteristics. The thinking and research reported here is recorded in roughly the same order. The report encompasses general background on the approach, and the results of some original research on approach taken in a meta-analysis, numerous illustrations of the approach, and the results of some original research on characteristics, (4) measuring the study findings on a common scale, and (5) analyzing the aggregation of findings and their relationship to the characteristics. The thinking can be read in at least three ways: as a textbook of methods of integrative analysis, as a record of some new ideas about integrative analysis, or as an apologia for meta-analysis. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208003

TM 800 675

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FINAL REPORT ON
GRANT #NIE-G-78-0148
PROJECT #8-0266

METHODS OF INTEGRATIVE ANALYSIS

to:

National Institute of Education

from:

Gene V Glass
Mary Lee Smith

Laboratory of Educational Research
University of Colorado

15 August 1980

INTEGRATION OF RESEARCH STUDIES:

Meta-analysis of Research

Gene V Glass
Laboratory of Educational Research
University of Colorado

Barry McGaw
Murdoch University
Perth, Western Australia

Karl R. White
Utah State University

Mary Lee Smith
Laboratory of Educational Research
University of Colorado

August, 1980

Ben-Adhem picked up a stone from beside the road. "It had written on it, 'Turn me over and read.' So he picked it up and looked at the other side. And there was written, 'Why do you seek more knowledge when you pay no heed to what you know already?'"

Shah (1968, p. 110)

"Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house."

Poincare (in La Science et l'Hypothese)

ABSTRACT

Integrative analysis, or what is coming to be known as meta-analysis, is the integration of the findings of many empirical research studies of a topic . For example, it might be undertaken to summarize the findings of fifty experiments on the effects of amphetamines on hyperactive pupils. Meta-analysis differs from traditional narrative forms of research reviewing in that it is more quantitative and statistical. Thus, the methods of meta-analysis are merely statistical methods, suitably adapted in many instances, that are applicable to the job of integrating findings from many studies.

A meta-analysis involves about a half-dozen steps: 1) defining the problem, 2) finding the research studies, 3) coding the study characteristics, 4) measuring the study findings on a common scale, and 5) analyzing the aggregation of findings and their relationship to the characteristics. The thinking and research reported here is recorded in roughly the same order. The report encompasses general background on the approach taken in a meta-analysis, numerous illustrations of the approach, and the results of some original research on statistical methods used in meta-analysis. The report can be read in at least three ways: as a textbook of methods of integrative analysis, as a record of some new ideas about integrative analysis, or as an apologia for meta-analysis.

PREFACE

This may be precisely the right time to write this book or precisely the wrong time. The reader ought not to assume that since this book lies before him that we came eventually to believe that the former was true. For we may have persevered in the face of ambivalence and written in spite of our doubts. Or we may have willfully written a book knowing that its time was not right. In fact, we wrote this book from necessity. Much of the work on which it is based has been supported for the past two years by a grant from the National Institute of Education. We are obliged now to file with the Institute some reasonable record of our efforts and their fruits. Propitious or not, this book will be written in this moment and not some later one. The reader who has ever struggled with writing a book will understand when we say that we now have that feeling that if we don't write it now, it will never get written.

Our subject is the methods of integrating empirical research. The problems we address lie at the center of a tiny revolution in the way social scientists and researchers attempt to extract knowledge from empirical inquiry and communicate it. The revolution was spawned by necessity. The findings of empirical research grew exponentially in the middle fifty years of the 20th century. Evidence — even the organized, analyzed and codified evidence of the archival journals -- multiplied beyond the ability of the unaided human mind to process it.

6

In the last ten years, scientists and methodologists have worried about the ways of synthesizing and organizing the findings of research into coherent patterns. We worried along with them, and we hope that our efforts helped clarify the problem if not solve small bits of it.

Because of our efforts and our colleagues' efforts and the efforts of a few dozen scholars around the world who have addressed the same problems, a person who starts out to review and integrate a body of research literature today has at his disposal some guidelines, examples and tricks that stand a pretty good chance of enriching his understanding of that literature. The methods of which we speak have now been applied a few hundred times, perhaps, and the experiences have been reported as being moderately satisfactory. (They have hardly escaped all criticism, but then what does?) In our minds, this counts as a hopeful beginning. But it is only a beginning - - and hereby lie our doubts about timing. A new field in its early stages should not have to contend with the conservative drag of a textbook the existence of which too often cuts off inquiry instead of stimulating it ("Well, if it's not in Grinch's Atlas of Organizational Dynamics, it must not be a problem"; or "Grinch says that's not so."). But, then, how does one weigh the disadvantages of a premature textbook against the disadvantages of no textbook at all? That question was ours, and we decided, "Better early than never."

CHAPTER ONE

THE PROBLEMS OF RESEARCH REVIEW AND INTEGRATION

The mathematician David Hilbert once said that the importance of a scientific work can be measured by the number of previous publications it makes superfluous to read. There is a hint of grouchiness and despair in Hilbert's complaint that scholars in all fields increasingly feel. What is one to make of the cornucopia of research literature? Can one make anything of it, or does one inevitably founder in the riches of empirical inquiry and sink to obselesence?

The house of social science research is sadly dilapidated. It is strewn among the scree of a hundred journals and lies about in the unsightly rubble of a million dissertations. Even it if cannot be built into a science, the rubble ought to be sifted and culled for whatever good there is in it.

Maccoby and Jacklin's (1974) review of research on the psychology of sex differences encompassed 1,600 works published before 1973. If one considers the literature on that topic sonce 1973 and realizes that many studies not focused specifically on sex differences may contain data on the question, then an estimated population of over 5,000 studies can be imagined. Dozens of educational problems could be named on which the available research literature numbers several hundred articles: ability grouping, reading instruction, programmed learning, instructional television, integration, etc. When Miller (see Smith, Glass and Miller, 1980) set out to determine the effects of drug therapy on psychological disorders, he found published reports of clinical experiments in such abundance (numbering literally thousands of studies) that he was forced to impose a sampling frame on the immense body of literature and take a survey sample of experiments! Social and behavioral research is a large and widely scattered enterprise. On problems of importance, it produces literally hundreds of studies in less than five years. The research techniques used, the measurements taken, the types of person studied -- each may vary in bewildering irregularity from one study to the next even though the topic is the same. The research enterprise in education and the social sciences is a rough-hewn, variegated undertaking of huge proportions. Determining what knowledge this enterprise has produced on some question is, itself, a genuinely important scholarly endeavor.

The style of research integration has been shaped by the size of the research literature. In the 1940's and '50's, a contributor to the Review of Educational Research or Psychological Bulletin might find one or two dozen studies on a topic. A narrative, rhetorical integration of so few studies was probably satisfactory. By the late 1960's, the research literature had swollen to gigantic proportions. Although scholars continued to integrate studies narratively, it was becoming clear that chronologically arranged verbal descriptions of research failed to portray the accumulated knowledge. Reviewers began to make crude classifications and measurements of the conditions and results of studies. Typically, studies were classified in contingency tables by type and by whether outcomes reached statistical significance. Integrating the research literature of the 1970's demands more sophisticated techniques of measurement and statistical analysis. The accumulated findings of dozens or even hundreds of studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than hundreds of data points in a single study could be so casually understood. Contemporary research reviewing ought to be undertaken in a style that is as technical and statistical as it is narrative and rhetorical. Toward this end, we suggested a name to make the needed approach distinctive. The desired approach was earlier referred to as the _meta-analysis_ of research (Glass, 1976). We have no stake in the use of this term; it sounds pretentious, but is only incidentally so. It was chosen to suggest the analysis of analyses, i.e., the statistical analysis of the findings of many individual analyses. The term _integrative analysis_ might serve as well, but meta-analysis has entered common parlance among some researchers fairly quickly and may become conventional. _Secondary-analysis_ is imprecise to the point of being misleading and should not be used interchangeably with these terms; it connotes an altogether different activity (Cook, 1974). Where a modification is needed to distinguish the meta-analysis of a body of studies from each of the studies individually, _primary research_ can be used to denote the latter.

Researchers have apparently thought little about the methodological and technical problems of research integration. Light and Smith (1971) first gave serious attention to these problems. Their paper is a careful treatment of the inadequacies of simple methods of research integration. Their proposed solution -- the cluster approach -- is in the spirit of the solution recommended here, but it is more conservative: ". . . little headway can be made by pooling the _words in the conclusions_ of a set of studies. Rather, progress will only come when we are able to pool, in a systematic manner, the original data from the studies." (Light and Smith, 1971, p. 443.) This assumption and the methods based on it probably discard far too many informative studies for which the data are no longer available, though the summary findings remain.

Gregg Jackson (1978), a sociologist, conducted what is perhaps the finest study yet of the practices and methods of research reviewers and synthesizers in the social sciences. He sampled at random 36 integrative reviews from the leading journals in education, psychology and sociology. The various features of method of each review were coded according to the categories of an extensive coding form that Jackson created. His conclusions:

a) Reviewers frequently fail to examine critically the evidence, methods and conclusions of previous _reviews_ on the same or similar topics. (Although 75 percent of the reviewers cited previous reviews, only 6 percent examined them critically.)

b) Reviewers often focus their discussion and analysis on only a part of the full set of studies they find, and the subset examined is seldom a representative sample nor is it clear how it (the subset) was chosen. (Only 3 percent of the reviewers appeared to have used existing indexes -- e.g., ERIC -- in their search; only 22 percent selected a fair sample of studies, in the judgment of Jackson's coders; and only 3 percent analyzed the full set of studies found.)

c) Reviewers frequently use crude and misleading representations of the findings of the studies. (About 15 percent of the reviewers classified studies according to whether their findings were "statistically significant," a practice which will be criticized in Chapter 5; frequently, reviewers report test-statistics (t, F, etc.) for one or more studies.

d) Reviewers sometimes fail to recognize that random sampling error can play a part in creating variable findings among studies.

e) Reviewers frequently fail to asses systematically possible relationships between the characteristics of the studies and the study findings. (Fewer than 10 percent of the reviewers studied whether the findings of the research were mediated by characteristics of the persons studied, the study context, the nature of the experimental intervention or the characteristics of the research design.) The lack of systematic examination of these relationships is important because reviewers frequently eliminate studies from consideration because of _a priori_ judgments that their findings are flawed by one or another study characteristic.

f) Reviewers usually report so little about their methods of reviewing that the reader cannot judge the validity of the conclusions.

Jackson also surveyed a small group of fewer than a dozen editors of review journals and executives of social science organizations in an attempt to determine which practices and standards prevail in their reviewing and integrating activities. He concluded that this survey was unproductive, but it was only unproductive of an articulated set of procedures and methods of study review and integration for the simple reason that such apparently do not exist. Jackson's small survey revealed clearly that the conception of research review and integration that prevails in the social and behavioral sciences is one in which the activity is viewed as a matter of largely private judgement, individual creativity and personal style. Indeed, it is and ought to be all of these to some degree; but if it is nothing but these, it is curiously inconsistant with the activity (viz., scientific research) it purports to illuminate.

Jackson (1978) went on in Chapter Six of his report to give a valuable list of guidelines for integrative reviewing that encompass such aspects of the process as selecting the topic, sampling studies, coding the characteristics of studies, analyzing, the data and interpreting the results. (Not coincidentally, guidelines for performing a primary research study could well be classified under the same headings.) Jackson devoted Chapter Four, "A New Alternative: Meta-Analysis" of his report to a description and critique of the approach that is the subject of this book.

Under the pressure of burgeoning research literatures, old and informal narrative techniques of research review and integration are breaking down. The fundamental problem is one of the mind's limitations and the magnitude of the task to which it is applied. The reviewer is even less able to absorb the sense of one hundred research studies than is an observer able to scan one hundred test scores and, without reliance on statistical methods, absorb the sense of their size and spread and correlations. Cooper and Rosenthal (1980) recently conducted an experiment in integrating research findings that illustrated these points. About forty persons (graduate students or more experienced) were randomly split into two groups. Subjects in both groups were given seven empirical studies on "sex differences in persistence" to review. Subjects in Group A were told:

> "Before drawing any final conclusions about the overall results of persistence studies, please take a moment to review each individual study. In generating a single conclusion from the independent studies, employ whatever criteria you would use if this exercise were being undertaken for a class term paper or a manuscript for publication."

Thus, Group A employed traditional, narrative techniques of integrating the findings of the seven studies. By contrast, Group B was instructed as follows:

> "Before drawing any final conclusions about the overall results of persistence studies, you are asked to perform a simple statistical procedure. The procedure is a way of combining the probabilities of independent studies. The purpose of the procedure is to generate a single probability level which relates to the likelihood of obtaining a set of studies displaying the observed results. This probability is interpreted just like that associated with a *t*- or *F*-statistic. For

example, assume the procedure produces a probability of .04. This would mean there are 4 chances in 100 that a set of studies showing these results were produced by chance. The procedure is called the Unweighted Stouffer method, and requires that you do the following:

1) Transfer the probabilities recorded earlier from each study to Column 1 of the Summary Sheet. [A summary sheet was provided each subject. The sheet contained the titles of the seven articles and columns for performing each step in the procedure.]

2) Since we are testing the hypothesis that females are more persistent than males, divide each probability in half (a probability of 1 becomes .5). If a study found men more persistent, attach a minus sign to it's probability. Place these numbers in Column 2. [It had been determined before hand that only two-tailed probabilities were reported.]

3) Use the Normal Deviations Table provided below and transform each probability in Column 2 into its associated Z-score. Place these values (with sign) in Column 3. If the probability is .5, the associated Z-score is zero (0).

4) Add the Z-scores in Column 3, keeping track of algebraic sign. Place this value at the bottom of Column 3.

5) Divide this number by the square root of the number of studies involved. In this case because N - 7, this number is 2.65. Thus, divide the sum of the Z-scores by 2.65. Place this number in the space below.

Z-SCORE FOR REVIEW_____________

6) Return to the Normal Deviations Table and identify the probability value associated with the Z-score for review. Place this number in the space below.

P-VALUE FOR REVIEW_____________

This probability tells how likely it is that a set of studies with these results could have been produced if there really were no relation between gender and persistence. The smaller the probability, the more likely it is that females and males differ in persistence, based on these studies." (cf. 1980, p. 445.)

Subjects in both Groups A and B rated their opinion of the strength of support for a conclusion of a relationship between sex and persistence in the seven studies. In fact, the combined results from the seven studies supported rejection of the null hypothesis of no difference at beyond the .02 level. The following frequencies were obtained:

	Group A Traditional Methods of Review		Group B Statistical Methods of Review	
Opinion (Is there a relationship?)	No.	%	No.	%
Definitely No	3	14%	1	5%
Probably No	13	59	5	26
Impossible to Say	5	23	8	42
Probably Yes	1	5	5	26
Definitely Yes	0	0	0	0
		100%		100%

The results are remarkable. Nearly 75 percent of the reviewers who relied on traditional narrative methods concluded that sex and persistence were not related; the comparable figure among the group using statistical methods of review was 31 percent -- rather strikingly different conclusions for equivalent groups trying to integrate only seven studies.

An issue of nearly equal importance concerns the magnitude of the relationship that the seven studies revealed. Again the reviewers in both groups were asked to rate their perception of the strength of the relationship.

	Group A Traditional Methods of Review		Group B Statistical Methods of Review	
Opinion (How large is the sex difference in persistence?)	No.	%	No.	%
None at all	4	18%	2	11%
Very small	12	55	6	32
Small	4	18	6	32
Moderate	2	9	4	21
Large	0	0	1	5
		100%		100%

The above data repeat the general findings apparaent in the previous table: persons using the two different methods of research integration formed quite different impressions about what the studies indicated. Cooper and Rosenthal examined these processes on a small collection of studies; the entire set of seven studies occupied a total of fewer than fifty journal pages. One can imagine how much more pronounced would be the difference between these two approaches with bodies of literature typical of the size of literatures that are increasingly being addressed with meta-analytic techniques. This difference will become more apparent to the reader who mends his way through the complex examples of research integration in the remainder of this book.

Consider another example of the contrasting conclusions arrived at through contrasting methods of review and integration. In a review of experiments on the effects of teachers' use of higher cognitive questions on students' achievement, Winne (1979) concluded that the former had no beneficial impact on the latter. A meta-analysis of virtually the same studies by Redfield and Rosseau (1980) revealed that on the average, students given higher cognitive level questions scored one-half standard deviation higher on achievement tests. Thus, informal and narrative techniques of review and integration discredited a finding that quantitative methods of integration showed to be consistent and large.

Narrative research reviews often make no attempt at rigorous definition and standardization of techniques for treating studies. Hence, impressions are subject to prejudice and stereotyping to a degree that would be unforgivable in primary research itself. Consider an instance encountered by Miller (1977) in his meta-analysis of experiments on the psychological benefits of drug therapy. At one point, attention focused on the question whether the combination of verbal psychotherapy and drug therapy was superior to the drug therapy alone. Three different traditional reviews completed within about five years of each other and based on largely the same literature arrived at the following conclusions:

> "The advantage for combined treatment is striking. . . a combination of treatments may represent more than an additive effect of two treatments -- a 'getting more for one's money' — there may also be some mutually facilitative interaction benefits for the combined treatments." (Luborsky, et.al., 1975, p. 1004).

> ". . . There is little difference between psychotherapy plus drug and drug therapy alone for hospitalized psychotic patients (but not for neurotic out-patients). The combination is, however, quite clearly superior to psychotherapy alone." (May, 1971, p. 513).

> "When all is said and done, the existing studies by no means permit firm conclusions as to the nature of the interaction between combined psychotherapy and medication. (Unlenhuth, Lipmen, & Covi, 1969, p. 611).

The disparity among these reviewers is not limited to their conclusions but extends even to their classification of individual experiments. Miller (1977) found five reviews (the three quoted above and two others) addressed specifically to the "psychotherapy - plus - drug" versus "drug therapy" issue. In Table 1. 1, the reviews, the studies reviewed and how they were classified are reported. Notice, for example, that Luborsky et. al., (1975) classified the Gorham study, the Cowden study and the King study as finding that "drug - plus - psychotherapy" was superior to "drug therapy" alone, whereas both Uhlenhuth (1969) and May (1971) in their reviews classified the same studies as showing no difference or a difference in the reverse order.

Obviously, different reviewers sometimes see things differently. The only way to force all reviewers to see the same thing is to demand a standardization of definitions and techniques of research integration. We don't suggest such; indeed, it would be ill-advised, since the little "reliability" that would be gained would probably be more than off-set by the creativity that would be staunched by uniformity.

It is not uniformity in research reviewing and integrating that is desirable, rather it is clarity, explicitness and openness — those properties that are characteristic of the scientific method more generally and which impart to inquiry its "objectivity" and trustworthiness.

Table 1.1

Summary of Findings of Five Reviews Comparing Drug
Plus Psychotherapy with Drug Therapy
(After Miller, 1977)

Reviewer	D + P>D	D>D + P or D + P = D
Group for the Advancement of Psychiatry (75)	King (58) Evangelikas (61) Klerman (74) Honigfeld (64)	May (64) Cowden (55,56)
Gilligan (65)	Evangelikas (61)	Cowden (56)
Uhlenhuth (69)	King (58) Evangelikas (61)	Cowden (56, 57) King (63) Honigfeld (64) Gorham (64) May (64)
Luborsky (75)	Gorham (64) Ilogarty (73) Cowden (56) King (63) Luborsky (54) Klerman (74)	King (60) May (65) Pascal (56) Evangelikas (61) Kroeger (67)
May (71)	Gorham (64)	King (63, 58) Gorham (64) Cowden (56) May (64) Evangelikas (61) Lorr (62)

It is often said of experimental research that is must be replicable to be scientific. Surely the true test of whether a finding is replicable is to replicate it; but as is observed ad naseum, studies never actually are replicated. Hence, the scientific attitude in research can not truly depend on replicability. Indeed, if one inquires more deeply into the question, one discovers that it is not replicability that is desirable in a scientific study, but the description of a study so that it could in theory be replicated, i.e., so that if one desired he could perform the same steps that led to the prior observations. Hence, to report a study so that it is "replicable" means to report it with such clarity and explicitness that a second investigator could follow the identical steps to the identical conclusion. Thereby, science is guaranteed to be "inter-subjective" rather than an endeavor subject to the whims and idiosyncracies of individual researchers. These values and standards are ingrained in the contemporary scientist's training; but too often he forgets his responsibility to the scientific method when he changes context slightly and seeks to integrate numerous empirical studies instead of perform a single primary study. Thus do reviews become idiosyncratic, authoritarian, subjective -- all those things that cut against the scientific grain.

The important point about the example in Table 1.1 is not that Uhlenhuth, Luborsky and May disagreed, but that they did not approach the problem of research integration with methods so explicit, unambiguous and operationally identified that any outside party could examine the same evidence and come to the same conclusion. By contrast, Miller (1977) approached the same research integration problem (viz., "drug - plus - psychotherapy" vs. "drug therapy") with an attitude like that of a researcher collecting and analyzing primary data: concepts must be defined and measured, measurements must be checked for reliability, evidence must not be excluded on arbitrary or ad hoc grounds, multiple observations inform on residual error, statistical methods are an important adjunct to raw perception. He found that the combined effect of drug and psychotherapy was approximately three-tenths standard deviations (on outcome measures of psychological well-being) greater than the isolated effect of drug therapy (see Chapter 8 in Smith, Glass and Miller, 1980).

CHAPTER TWO

META-ANALYSIS OF RESEARCH

Primary analysis is the original analysis of data in a research study. It is what one typically imagines as the application of statistical methods.

Secondary analysis is the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data. Secondary analysis is an important feature of the research and evaluation enterprise. Tom Cook (1974) at Northwestern University has written about its purposes and methods. Some of our best methodologists have pursued secondary analyses in such grand style that its importance has eclipsed that of the primary analysis.

But our topic is what we have come to call -- not for want of a less imposing name -- meta-analysis of research. In 1976, one of us defined it thus:

> "Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature." (Glass, 1976, p. 3).

And again, two years later:

> "The accumulated findings of dozens or even hundreds of studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than hundreds of data points in a single study could

> be so casually understood. Contemporary research reviewing ought to be undertaken in a style more technical and statistical than narrative and rhetorical. Toward this end, I have suggested a name to make the needed approach distinctive; I referred to this approach as the meta-analysis of research (Glass, 1976). I have no stake in the use of this term; it sounds pretentious, but is only incidentally so. It was chosen to suggest the analysis of analyses, i.e., the statistical analysis of the findings of many individual analyses." (Glass, 1978, p. 352).

And two years later still:

> "The approach to research integration referred to as 'meta-analysis' is nothing more than the attitude of data analysis applied to quantitative summaries of individual experiments. By recording the properties of studies and their findings in quantitative terms, the meta-analysis of research invites one who would integrate numerous and diverse findings to apply the full power of statistical methods to the task. Thus it is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis." (Glass, 1980, p. 2).

The essential character of meta-analysis is that it is the statistical analysis of the summary findings of many empirical studies.

Meta-Analysis Is Quantitative

Meta-analysis is quantitative. It is undeniably quantitative; and by and large it uses numbers and statistical methods in a practical way, namely, for organizing and extracting information from large masses of data that are nearly incomprehensible by other means. Numerosity creates many of the problems of research synthesis; naturally, numerical methods are employed in their solution.

Meta-Analysis Does Not Prejudge Research Findings in Terms of Research Quality

The findings of studies are not judged a priori or by arbitrary and non-empirical criteria of research quality. In this respect, meta-analysis differs greatly from other approaches to research integration. Typical narrative reviews attempt to deal with multiplicity by arbitrary exclusion. The dissertation literature is excluded because it may be believed that any worthwhile study would have been published. Huge numbers of studies are excluded on methodological grounds: poor design, bad measurement, badly implemented treatment, and the like. Yet, evidence is never given to support these arbitrary exclusions.

An important part of every meta-analysis with which we have been associated has been the recording of methodological weaknesses in the original studies and the examination of their relationship to study findings. Thus, the influence of study quality on findings has been regarded as an empirical _a posteriori_ question, not an _a priori_ matter of opinion or judgment used to exclude large numbers of studies from consideration.

Meta-Analysis Seeks General Conclusions

The most common criticism of meta-analysis is that it is illogical because it mixes findings from studies that are not the same; it mixes apples and oranges. Implicit in this concern is the belief that only studies that are the _same_ in certain respects can be aggregated. The claim that only studies which are the same in _all_ respects can be compared is self-contradictory; there is no need to compare them since they would obviously have the same findings within statistical error. The only studies which need to be synthesized or integrated are _different_ studies. Generalizations will necessarily entail ignoring some distinctions that can be made among studies. Good generalizations will be arrived at by ignoring only those distinctions that make no important difference. But ignore we must; knowledge itself is possible only through the orderly discarding of information.

Yet it is intuitively clear that some differences among studies are so large or critical that no one is interested in their integration. What, for example, is to be made of study #1 which demonstrates the effectiveness of disulfiram in the treatment of alcoholism and study #2 which demonstrates the benefits of motorcycle helmet laws? Not much, I suppose. But it hardly follows that the integration of study #1 on lysergide treatment of alcoholism and study #2 on "controlled drinking" is meaningless; one is understandably concerned with which treatment has a greater cure rate. Is the essential difference between the two examples that in the former case the problems addressed by the studies are different but the problem is the same in the latter example? "Problem" is no better defined than "study" or "findings," and invoking the word clarifies little. It is easy to imagine the Secretary for Health comparing fifty studies on alcoholism treatment with fifty studies on drug addiction treatment or a hundred studies on the treatment of obesity. If the two former groups of studies are negative and the latter is positive, the Secretary may decide to fund only obesity treatment centers. From the Secretary's point of view, the problem is public health, not simply alcoholism or drug addiction treatment.

There exists another respect in which it is inconsistent to criticize meta-analysis as meaningless because it mixes apples and oranges.

Data analyses of primary research are traditionally performed by lumping together (averaging or otherwise aggregating in analyses of variance, *t*-tests and whatever) data from different persons. These persons are as different and as much like apples and oranges in their way as studies are different from each other. Yet to object to pooling the findings of studies 1, 2, . . ., 10 and see nothing at all objectionable in pooling the results from persons 1, 2, . . ., 100 is inconsistent. Now one might think that the two kinds of aggregating identified are qualitatively different; but it would remain to be specified exactly how they are different and why it matters, which would necessarily entail presenting empirical evidence to demonstrate that studies using different populations, measuring instruments, data analyses, etc. are fundamentally incommensurable. The ironic dilemma posed here is that such an empirical demonstration would be of itself an analysis of exactly the type which we have referred to as a "meta-analysis."

Meta-analysis is aimed at generalization and practical simplicity. It aims to derive a useful generalization that does not do violence to a more useful contingent or interactive conclusion. The world runs on generalizations and marginal utilities. They represent synthesis; science runs on analysis. Therein lie many of the difficulties that scientists and men of practical affairs encounter when they meet.

Our approach, meta-analysis, has been misunderstood -- a circumstance for which we must accept that share of the responsibility due us. It has been characterized by some as "averaging effect sizes," which is a little like characterizing analysis of variance as "adding and multiplying." The sine qua non of what we call meta-analysis is the application of research methods to the characteristics and findings of research studies. By "research methods" is meant such considerations as are normally addressed in conceptualizing, designing and analyzing empirical research: problem selection, hypothesis formulation, definition and measurement of constructs and variables, sampling, data analysis (see Kerlinger, 1964, or many others).

The methods of meta-analysis have much in common with those of survey research, for in fact, research review and integration is a process of surveying and analyzing in quantitative ways large collectivities. Many of the issues faced in a meta-analysis are akin to the problems addressed in survey design and analysis (cf. Kish, 1965). The similarity between the two should not be taken as implying that meta-analysis shares with survey research the latter's limitations as regards the analysis of causal claims. Survey research continues to struggle with the problems of unknown third variables and ambiguous direction of causality. Meta-analysis, on the other hand, through no great accomplishment of its' own, may very well be applied to the findings of a literature of controlled experimental studies, each of his has a valid claim on a causal conclusion.

We do not wish to imply that a clear break can be discerned between earlier methods of research integration and meta-analysis. In fact, under the pressure of numbers, research reviewers have gradually of necessity adopted increasingly rigorous and quantitative methods of study integration in the past thirty years. For example, Underwood (1957) found 16 experiments on the link between memory and interference when he attempted to integrate the existing research. The standard designs and the near standard measurements common to the studies suggested a more quantitative amalgamation of the evidence than was typical in research reviewing at the time. By graphing the number of lists of items to be recalled in these experiments against the percent correct recall on the last list, Underwood obtained an orderly and convincing pattern describing the relationship (see Figure 2.1). By portraying multiple findings quantitatively and aggregating across some potentially irrelevant distinctions (e.g., lists of geometric forms vs. nonsense syllables; paired-associate vs. serial presentation, long lists vs. short lists), Underwood discovered a convincing and important finding not apparent in the disparate constituent studies. This is the essence of the meta analysis approach.

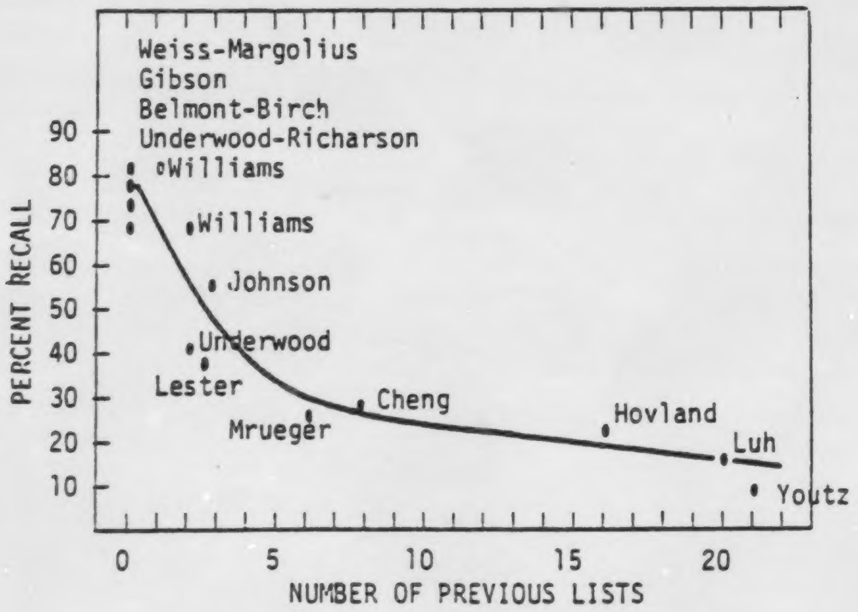


Figure 2.1 Recall as a function of previous lists learned as determined from a number of studies.

Readers of review journals (e.g., Psychological Bulletin, Review of Educational Research, American Sociological Review) have become familiar with increasingly more elaborate forms of research integration. Long lists of coded descriptions of research literatures have become common. Contingency tables showing proportions of "significant results" under various conditions are more and more a standard feature of integrative reviews. These developments were required by the complexity of the reviewing task, and they are in the spirit of the methods we present here. We hope to have advanced these methods by appropriately increasing the quantification and analysis of the task so that the full value of modern statistical methods is realized.

Rosenthal (1976) integrated the findings of several hundred studies of the experimenter expectancy effect in behavioral research. The techniques he used and his discussion of methodology were remarkably like those presented in Glass (1976), though the two efforts (borne of similar necessities) proceeded independently. In the five years since our work has been publicized, the methods developed and recommended have been applied repeatedly and in diverse areas: treatment of stuttering (Andrews, 1979), modern vs. traditional math instruction (Athappilly, 1980), "process oriented" science instruction (Bredderman, 1979), mainstreaming of special education students (Carlberg, 1979), neuropsychological assessment of children (Davidson, 1978), "inquiry oriented" science teaching (El-Nemr, 1979), transcendental meditation (Ferguson, 1980), teaching style and pupil achievement

(Glass et al., 1977; Gage, 1978), social-psychological environments and learning (Haertel, Walberg and Haertel, 1979), sex differences in decoding verbal cues (Hall, 1978), individualized mathematics instruction (Hartley, 1977), effects of television on social behavior (Hearold, 1979), validity of employment tests (Hunter, Schmidt, and Hunter, 1979), home environment and learning (Iverson and Walberg, 1979), psycho-linguistic training (Kavale, 1979), treatment of hyperactivitiy (Kavale, 1980), racial desegregation and academic achievement (Krol, 1979), personalized college-level instruction (Kulik, Kulik and Cohen, 1979), advance organizers (Luiten, Ames and Ackerson, 1979), drug therapy and psychological disorders (Miller, 1977; and Smith, Glass and Miller, 1980), test validity in personnel selection (Pearlman, 1979), teachers' questioning style (Redfield and Rousseau, 1979), psychotherapy and medical utilization (Schlesinger, Mumford and Glass, 1978), psychotherapy and recovery from medical crisis (Schlesinger, Mumford and Glass, 1979), aesthetics education and basic skills (Smith, 1980), sex-bias in counseling and psychotherapy (Smith, 1980), class-size and affective outcomes (Smith, and Glass, 1979), psychotherapy outcomes (Smith and Glass, 1977), motivation and achievement (Uguroglu and Walberg, 1978), socioeconomic status and academic achievement (1976), relationship between attitude and achievement (Willson, 1980), patient education programs in medicine (Posavac, 1980), correlation of auditory perceptual skill and reading (Kavole, 1980), diagnostic/remedial instruction and science learning (Yeany and Miller, 1980), treatment of migraine and tension headache (Blanchard, Andrasik, Ahles, Teders and O'Keefe, 1980), effects of direct versus open instruction (Peterson, 1978).

Illustrations of Meta-Analysis

Meta-analysis has been misunderstood and criticized, the criticisms often gathering their force from the misunderstandings. But the objections raised to meta-analysis are the subject of the final chapter. In the remainder of this chapter, we wish instead to elaborate on the verbal characterization of meta-analysis by describing briefly several applications of the method.

Psychotherapy and Asthma. Twelve studies were located that tested the effects of psychotherapy on asthma. Eleven studies used treatment and control group designs; two designs were pretest versus posttest.

The summary of the data and findings appear as Table 2.1 which offers the following items of information about each study: a) Author(s); b) type of therapy; c) average age of subjects; d) number of hours of therapy given; e) the nature of the control group (no treatment, relaxation therapy, medical treatment); f) the number of weeks elapsing between the end of therapy and measurement of the outcome variable; g) the nature of the dependent (outcome) variable; h) the effect (ES) achieved in the study, the treatment mean minus the control mean divided by the control group standard deviation, viz.,

$$ES = \frac{\overline{X}_{psy} - \overline{X}\text{control}}{\sigma \text{ control}}$$

Table 2.1

Findings of 11 Studies of Psychological Treatment of Asthma

Study	Therapy Type	Age	Hours of Therapy	Control Group	Follow-up Time (weeks)	Dependent Variable	ES
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Moore (1965)	Reciprocal Inhibition	21 1/2 adults 1/2 children	4	Relax Training	0	Lung functioning No. asthma attacks	1.41 .88
Sclare, et al.	Psycho-dynamic	30 (19-42)	28	Physical Treatment	0	Remission of symptoms	.66
Yorkstun et al.	Verbal Desensitization	42	3	Relax Training	0	Lung functioning	1.00
					96	Psychiatrist's rating of improvment	1.00
					96	Use of drugs.	1.52
Maher-Loughnan, et al. (1962)	Hypno-Therapy	25	20	No treatment	0	Symptoms, wheezing	.64
Citron, K. M. (1968)	Hypnotherapy	30	12	Relax Training	0	Symptoms, wheezing	.52
Groen & Pelser (1960)	Psycho-dynamic (group)	45	50	Medical treatment	24	Rated Improvement	1.36
Barendregt (1957)	Eclectic (4 dynamic)	42	100	Medical treatment	0	Increased hostility, decreased "oppression & damage; Rorschach	.57
Ago, et al. (1976)	Eclectic (4-somatic 4-therapy)	34	20	Medical treatment	120	Remission of asthma symptoms	1.51

Table 2.1 continued

Findings of 11 Studies of Psychological Treatment of Asthma

Study	Therapy Type	Age	Hours of Therapy	Control Group	Follow-up Time (weeks)	Dependent Variable	ES
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Kahn (1977)	Counter-conditioning	12	15	No treatment	32	Use of drugs & medication	.29
					32	Hospitalization	.19
					32	Asthma attacks	.24
Kahn, *et al.* (1973)	Counter-conditioning	11	15	Medical treatment	40	No. of ER visits	.76
					40	Amount of drugs & medication	1.11
					40	No. of asthma attacks (one hospitalization in control group, none in therp.)	.66
Alexander *et al.*	Jacobson relaxation training	12	3	No treatment	0	Pulmonary functioning (peak expiratory flow)	.82
McLean, A. F. (1965)	Hypnotherapy	13	6	None (pretest vs. posttest)	12	Wheezing Score	1.23
Arnoff, G. M. *et al.*	Hypnotherapy	10	1/2	None (pretest vs. posttest)	0	Forced lung capacity	0.71
						Peak air flow rate	0.67
						Dyspnea	1.25

The overall (i.e., summed across all studies) measure of impact of psychotherapy on asthama is depicted in Figure 2.2.

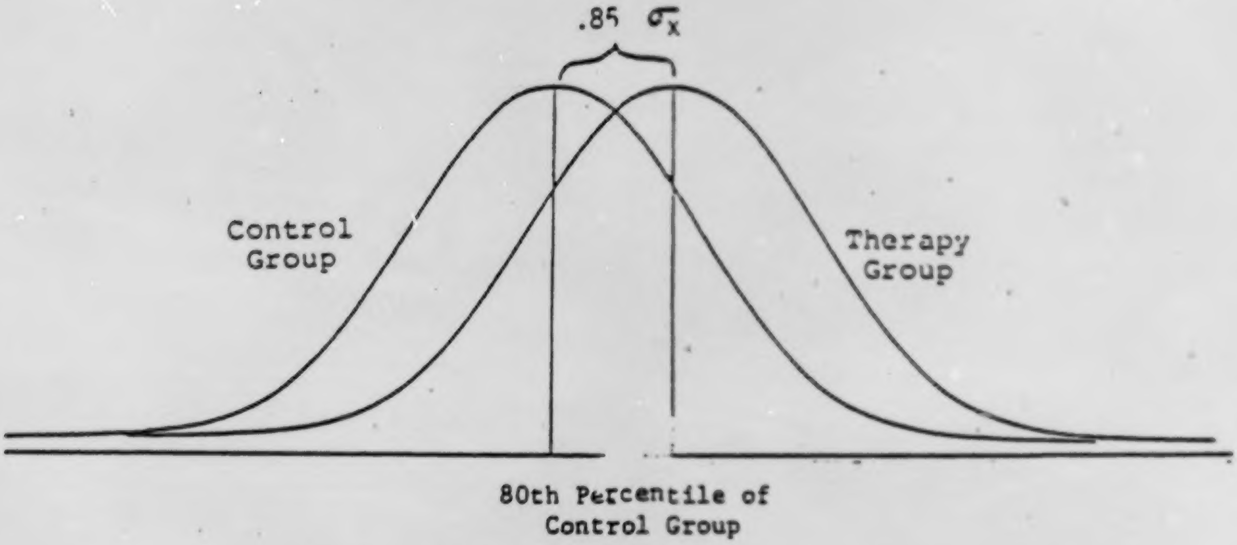


Figure 2.2 Average effect of psychotherapy on asthma outcome measures across 13 studies which included 23 outcome variables.

The average effect comparing therapy and control groups was .85 σ_x , i.e., the average subject who received psychotherapy was at .85 standard deviations above the mean of the untreated controls. (The standard deviation of the 23 effect size measures is $\sigma_{ES} = .390$; thus, the 95% confidence interval of the true average ES is $.85 \pm \frac{1.96\ (.390)}{\sqrt{23}} = (.69, 1.01,)$. It follows that the average therapy subject exceeds 80% of the untreated controls on the aggregate outcome variables.

There were six outcome measures in the thirteen studies that assessed the use of medical services: use of medicine, hospitalization, emergency room visits. The average effect size for these six outcomes was $\overline{ES} = .73$. The two summary effect sizes — .85 for all outcomes and .73 for direct medical services -- compare favorably with the effects of psychotherapy on outcomes such as fear, anxiety, and self-esteem.

The relationship between the effects of psychotherapy and some features of the therapy and the patients is examined in Tables 2.2 through 2.5.

Therapy Type: The average effect sizes by type of therapy are as follows:

Table 2.2

Type of Therapy

	Behavioral	Psychodynamic	Hypnotherapy	Relaxation
n:	12	4	6	1
$\overline{ES}$:	.80	1.03	.84	.82
σ_{ES}:	.42	.41	.79	0

The differences among the effects of different types of therapy are not large, and in no case do they reach conventional levels of statistical significance.

Age of Patient: The distribution of patients' ages (averaged within each study) is as follows:

Table 2.3

Age

	10-15	16-20	21-25	26-30	31-35	36-40	41-45
Frequency:	5	0	2	2	1	0	3

The linear correlation between age of patients (at the study level) and ES is +.40, which is reasonably statistically significant (standard error of $r \cong .21$).

Hours of Therapy: The distribution of duration of therapy in hours for the 13 experiments is as follows:

Table 2.4

Hours of Therapy

	1-5	6-10	11-20	21-50	51-100	
Freq.	9	1	10	2	1	$\bar{X}. = 21.3$
$\overline{ES}$.:	1.03	1.23	.64	1.01	.57	

The linear correlation of "hrs. of therapy" and ES across the 23 outcome measures is - .15, not significantly different from zero.

Follow-up Time: The follow-up times for measurement of effects for the 23 outcome measures were distributed as follows:

Table 2.5

Weeks Post Therapy

	0	12	24	32	40	96	120	
Frequency:	11	1	1	4	3	2	1	$\bar{X}. = 25.9$
$\overline{ES}$.:	.81	1.23	1.36	.30	.84	1.26	1.51	

The linear correlation of "weeks post therapy" and $\overline{ES}$ is .34, not significantly different from zero at any respectable significance level.

Psychotherapy (primarily behavioral therapies and hypnotherapy) shows impressively large effects on ameliorating the effects of asthma. The effects are even substantial on the reduction of utilization of direct medical services, showing a reduction in utilization such that only 23 percent of the therapy subjects used as many medical services as half the control subjects. It is important to note, in this regard, that in 5 of the 11 experimental vs. control group studies, the control group received medical treatment that was not given to the psychotherapy group.

Psychotherapy and Alcoholism. In Table 2.6 appear data from 15 experiments on the effects of psychotherapy on alcoholism. In successive columns appear the following information about each study:

a) The investigator(s) and year of the study;

b) The type of therapy administered (e.g., behavioral modification, eclectic, psychodynamic);

c) The number of hours of therapy administered;

d) The number of months after therapy at which outcomes were measured;

e) A definition of "success" for the outcome measure;

f) The percentage of "successes" in the therapy group;

g) The percentage of "successes" in the control group;

h) The differential success, Δ: f) minus g), above.

Summary tablulations of a few characteristics of the studies in Table 2.6 are presented below:

Types of therapy: 11 studies used non-behavioral therapy.

9 studies used behavioral therapy.

Distribution of hours of treatment:

	Hours						
	1-10	11-20	21-30	31-40	41-50	51-60	61-75
Frequency:	4	9	3	0	2	1	1

Distribution of follow-up times:

	Months Past Therapy				
	0	1-3	4-6	7-9	10-12
Frequency:	4	0	10	1	5

Of course, interest centers primarily on the outcome measures. There exist two approaches to summarize the outcomes: 1) the data can be pooled across all studies to calculate aggregate "success" rates, or 2) the "success" rates can be averaged across the 15 studies. The first method gives a study an importance in the aggregate which is proportional to its sample size, which could be desirable in some instances but probably isn't in this instance. The second method weights each study equally, in effect.

By the first method of aggregation, one finds 651 patients treated with psychotherapy with 269 reported as "successes" for a success rate of 41 percent. The comparable figures for the control condition are 638 cases, 222 "successes" for a "success" rate of 33 percent. The 41 percent vs. 33 percent difference is not very impressive; but it may not be very fair. Note that a few studies like Gallant (1971) and McCance and McCance (1969) carry unreasonably large weight in determining these aggregates because between them they account for nearly half of all the therapy cases.

Averaging success rates across studies seems preferable. Doing so yields "success" rates of 51 percent and 33 percent for psychotherapy and control conditions, respectively. These figures are probably more defensible than the 41 percent vs. 33 percent figures. Even, so, a "success" rate of 33 percent for untreated controls is unusual and indicates that the experiments were probably conducted under favorable circumstances with other than chronic

Table 2.6

Results of Outcome Studies on Psychological Treatment of Alcoholism

						Outcomes		
a)		Type of Therapy b)	Hrs. of Therapy c)	Mos. post-therp. for follow-up d)	Type of Outcome e)	Percent Success in Therp. f)	Percent Success in control g)	Δ %
Vogler *et al.*	'70	Beh. Mod.	15	8	Not relapsed into alcoholism	$\frac{14}{25} = 56\%$;	$\frac{5}{12} = 42\%$	6%
Cadogan	'73	Eclectic	18	6	Abstinence	$\frac{18}{20} = 90$	$\frac{4}{20} = 20$	70
Clancy *et al.*	'69	Beh. Mod.	4	12	Abstinence	$\frac{6}{25} = 24$	$\frac{3}{17} = 18$	6
Gallant	'71	Eclectic	50	0	Sobriety	$\frac{17}{140} = 12$	$\frac{3}{70} = 4$	8
Gallant *et al.*		Psychodynam.	50	0	Sobriety	$\frac{2}{21} = 10$	$\frac{1}{21} = 5$	5
Gallant *et al.*	'68	Psychodynam.	60	0	Abstinence or nearly so	$\frac{7}{10} = 70$	$\frac{1}{9} = 11$	59
Hunt & Azrin	'73	Eclectic	75	0	Abstinence	$\frac{7}{8} = 88$	$\frac{1}{8} = 13$	74
McCance &		Psychodynam.	12	6	Abstinence or nearly so	$\frac{20}{31} = 65$	$\frac{23}{51} = 45$	20
McCance	'69	Psychodynam.	12	12	Abstinence or nearly so	$\frac{13}{30} = 43$	$\frac{23}{49} = 47$	- 4
McCance	'69	Beh. Mod.	6	6	Abstinence or nearly so	$\frac{24}{45} = 53$	$\frac{23}{51} = 45$	8
McCance	'69	Beh. Mod.	6	12	Abstinence or nearly so	$\frac{24}{45} = 53$	$\frac{23}{49} = 47$	6

Table 2.6 (continued)

a)	Type of Therapy b)	Hrs. of Therapy c)	Mos. post-therp. for follow-up d)	Type of Outcome e)	Percent Success in Therp. f)	Percent Success in control g)	Δ %
Kissin et al. '70	Psychodynam.	20	6	Abstinence or nearly so	$\frac{22}{62} = 35$	$\frac{2}{44} = 5$	30
Kissin et al. '70	Psychodynam.	20	6	Abstinence or nearly so	$\frac{5}{33} = 15$	$\frac{2}{41} = 5$	10
Sobell & Sobell '73	Beh. Mod.	25	6	Full or part-time employ.	$\frac{21}{35} = 60$	$\frac{14}{35} = 40$	20
Sobell & Sobell	Beh. Mod.	25	12	Full or part-time employ.	$\frac{21}{35} = 60$	$\frac{16}{35} = 46$	14
Levinson & Sereny '69	Eclectic	30	12	Slight or much improv.	$\frac{15}{26} = 58$	$\frac{17}{27} = 63$	- 5
Newton & Stein '72	Eclectic	15	6	Not readmitted to hosp. for alcohol.	$\frac{10}{15} = 67$	$\frac{11}{16} = 69$	- 2
Newton & Stein '72	Implosive	15	6	" " " "	$\frac{7}{15} = 47$	$\frac{11}{16} = 69$	- 22
Ashem & Donner '68	Beh. Mod.	5	6	Sobriety	$\frac{6}{15} = 40$	$\frac{0}{8} = 0$	40
Storm & Cutler '70	Sys. desen.	12	6	Some or marked improv.	$\frac{10}{15} = 67$	$\frac{39}{62} = 63$	4

alcoholics. But even if the 33 percent base-rate figure is unrealistic, the 18 percent gap between treatment and control groups is not. One can conclude that on the average 20 hours of psychotherapy produces 18 "successes" (sobriety 6 months after therapy) out of every 100 persons treated.

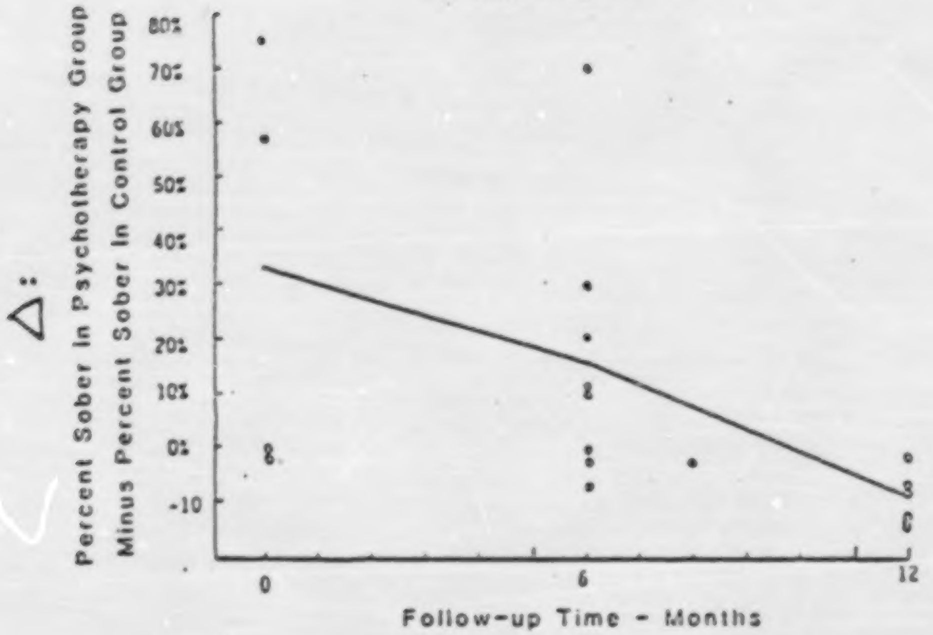
The percentage "success" rates can be transformed into a metric measure of effect by means of the probit transformation (Glass, 1978). A discrepancy of 51 percent to 33 percent corresponds to a metric measure of effect of +.96 standard deviation units. Expression of the effect in this way will permit comparison of the effects across problem areas such as alcoholism, asthma, and surgery.

The relationship of the differential success rate to follow-up time and amount of therapy was also studied (see figure below.) The difference in percentages of "successes" between treatment and control groups diminished across follow-up intervals. Immediately after therapy, there were 37 percent more successes in the therapy group than the control group; at six months after therapy this difference dropped to 25 percent; at twelve months it was 3 percent, i.e., the rate of sobriety is virtually the same in the treatment and control groups, the treated patients having relapsed. Apparently, for the benefits of the therapy to be sustained, it must be readministered at periodic intervals.

Finally, the correlation across the 15 studies between the number of hours of therapy and the differential "success" rate was positive and reasonably large: +.49. More therapy was better than less.

DECAY OF TREATMENT EFFECT

ALCOHOLISM



Solid Line connects averages at 0, 6, and 12 months.

School Class-Size and Achievement

The literature on school class-size and its relationship to achievement has lain about for many years. Some of the first empirical research in education, that of Joseph M. Rice in the 1890's, examined the association between class-size and learning. In graduate school in the 1960's, I was taught that the two were unrelated and there was little point pursuing the matter. A faint aroma of Chippendale (unwieldy and antique) still clung to the topic when in 1977 a friend at the Far West Laboratory, Leonard Cahen, suggested that we apply to the class-size literature the techniques we had developed for integrating outcome experiments in psychotherapy. The $8,000 contract he dangled before us, made the problem seem worthwhile.

The literature on class-size and achievement had been reviewed repeatedly. The reviewers disagreed wildly. One could document this confusion; it would be simple to quote reviewer X claiming that large classes are better, reviewer Y that small classes are better, and reviewer Z that neither is better. But to do so would only embarrass others and add nothing to one's appreciation of the complexity of the research. The problems with previous reviews of the class-size literature are several: (1) literature searches were haphazard and often overly selective; dissertations were avoided, as a rule, and few reviewers sought out large archives of pertinent data; (2) reviews were typically narrative and discursive; the multiplicity of findings could not be absorbed without quantitative methods of reviewing; (3) reviewers that attempted quantitative integration of findings made several mistakes — they used crude classifications of class-sizes; and (4) they took statistical significance of differences far too seriously.

Our search for class-size studies was carried out in three places: (1) document retrieval and abstracting resources; (2) previous reviews of the class-size literature and (3) the bibliographies of studies once found. The ERIC system and Dissertation Abstracts were searched completely on the key words "size," "class size," and "tutoring." The dissertation literature was covered as far back as 1900, and the fugitive educational research literature was covered from the mid 1960's to 1978. Of the many hundreds of doctoral dissertations scanned in Dissertation Abstracts, about thirty micro-film copies were purchased. A dozen dissertations were eventually incorporated. The journal literature on class-size was located in the traditional way; one or two current reviews of the research were found, the articles cited were located, and the articles cited in these articles were located in turn. About 300 documents were obtained and read. One hundred-fifty of them were found to contain no usable data, i.e., no data whatsoever were reported on the comparison of small- and large-class achievement. About 70 studies examined the relationship of class-size to non-achievement outcomes and classroom process variables. Approximately 80 studies on the class-size and achievement relationship were included in the meta-analysis.

It is difficult to estimate what portion of the existing literature was captured by this search. Even though 80 studies exceeded by 50 percent the most extensive reviews published to that time, perhaps less than half of all studies that exist on the topic were found. Some studies (credited to school districts) could not be located even after several phone calls and letters. Other studies were surely missed because of odd or nondescript titles. Fortunately, the ERIC system uses key words based on the contents of a paper and not titles alone. Several studies found in the journal literauture by branching off existing bibliographies had neither "size" nor "class-size" in the title, evidence enough that several studies were missed because their titles lacked the key words. Another complication concerns the use of class-size as an incidental variable in studies focused on other issues. There are probably many such studies, and only a few of the most visible ones were located.

The research on class-size and its relationship to achievement evolved through four stages: the pre-experimental era (1895-1920); the efficiency era (1920-1940); the large-group technology era (1950-1970); and the individualization era (1970-present). The boundaries of the eras are not impenetrable, and even today an atavistic throwback to the 19th century will appear in a doctoral thesis. At each new stage, the sophistication of research methodology increased, and the question of class-size and its effect on achievement was examined with different motives. One discerns in the narration accompanying the numbers the cult of efficiency of the early part of this century, the rising birth rate of the post-war '40's, the advent of teaching technology in the '60's, and most recently the teacher labor movement and declining enrollments. What was said about the data changed as new interpretations served emerging purposes, even when the data changed little themselves.

The meta-analysis was to determine what the available research revealed about the relationship of class-size to achievement. Drawing boundaries around this topic was simple compared to the difficulties encountered in defining psychotherapy, for example (Smith and Glass, 1977). Conventional definitions of achievement seem scarcely to have changed over eighty years; and class-size is relatively easily described and measured.

The quantification of characteristics of studies permitted the eventual statistical description of how properties of studies affect the principal findings. Such questions can be addressed as "How does the class size and achievement relationship vary as a function of age of pupils?" or "How does it vary between reading and math instruction?" The first step was to identify those properties of studies that might interact with the relationship between class-size and achievement. There is no systematic logical procedure for taking this step. One simply reads a few studies from the literature of interest, talks with experts, and then guesses; modifications can always be made later if needed. About 25 specific items were coded for each study. Some were more useful than others; several items were seldom reported in the studies. A coding sheet was devised onto which the information about each study was transcribed. A single study might fill several coding sheets, depending on how many different class sizes were compared, how many different achievement tests were reported separately for different ages or IQs, and so forth.

The major items of the coding sheet were as follows: (1) year of publication; (2) publication source (book, thesis, journal); (3) subject taught (reading, math, etc.); (4) duration of instruction (number of weeks); (5) number of pupils in the study (different from class-size since there might be many classes); (6) number of teachers in the study; (7) pupil ability; (8) pupil ages; (9) types of experimental control (random assignments, matching, etc.); (10) achievement measurement (standardized test, ad hoc test, etc.); (11) quantification of outcomes (gain scores, ANCOVA adjustment, etc.)

A simple statistic was desired that described the relationship between class-size and achievement as determined by a study. No matter how many class-sizes are compared, the data can be reduced to some number of pairs, a smaller class against a larger class. Certain differences in the findings must be attended to if the findings are later to be integrated. The most obvious difference is the scale properties of the achievement measure. Measurement scales can be standardized by dividing mean differences in achievement by the within group standard deviation (a method that is complete and discards no information at all under the assumption of normal distributions). The eventual measure of relationship seems straight-forward and unobjectionable:

$$\Delta_{S-L} = \frac{\overline{X}_S - \overline{X}_L}{\hat{\sigma}} , \qquad (2.1)$$

where:

$\overline{X}_S$ is the estimated mean achievement of the smaller class which contains S pupils.

$\overline{X}_L$ is the estimated mean achievement of the larger class which contains L pupils: and

$\hat{\sigma}$ is the estimated within-class standard deviation, assumed to be homogeneous across the two classes.

As a first approximation to studying the class-size and achievement relationship, it is considered irrelevent that the particular types of achievement that lie behind the variable X are quite different knowledges and skills measured in quite different ways. Reports of research frequently omit such basic descriptive measures as means and standard deviations. This omission frequently complicated the calculation of Δ_{S-L}, but seldom obviated it. Transformations of commonly reported statistics (*t*, *F*, etc.) into Δ's were derived (Glass, 1978).

In all, 77 different studies were read, coded, and analyzed. These studies yielded a total of 725 Δ's. The comparisons are based on data from a total of nearly 900,000 pupils spanning 70 years research in more than a dozen countries. In Table 2.7 appears the frequency distribution of Δ's by year in which the study appeared. It is clear from Table 2.7 that class-size research was an active early topic in educational research, was largely abandoned for 30 years after 1930, and has been resurrected in the last 15 years. In Table 2.8, the comparisons are tabulated by the type of assignment of pupils to the different size classes. Each of the first three types of assignment represents reasonably good attempts at eliminating gross inadequacies in design; these three conditions account for slightly more than half of all the comparisons. Even though half of the comparisons involved comparing naturally constituted and non-equivalent large and small classes, some of these were based on *ex post facto* statistical adjustments for pre-existing differences. So the data are not half worthless; indeed, whether the experimental inadequacies influenced the findings is an empirical question — rather than an *a priori* judgment — which was examined in the data analyses. In Table 2.9 appears the joint distribution of smaller and larger class-sizes on which the 725 Δ's are based. For example, six Δ's derive from comparisons of group sizes 1 and 3. (The table contains only 550 entries instead of 725, since comparisons would not be recorded in this tabulation if S and L were contained within the same broad category (e.g., if $S = 18$ and $L = 22$.)

Table 2.7

Class-Size Comparisons (Δ) by Year of Study

Year	No. of Δ's	%	Cumulative %
1900-1909	22	3.0%	3.0%
1910-1919	184	25.4%	28.4%
1920-1929	138	19.0%	47.4%
1930-1939	47	6.5%	53.9%
1940-1949	1	0.0%	53.9%
1950-1959	62	8.6%	62.5%
1960-1969	150	20.8%	83.3%
1970-1979	121	16.7%	100.0%
	725	100.0%	

Table 2.8

Class-Comparisons (Δ) by Assignment of Pupils to the Small and Large Classes

Type of Assignment	No. of Δ's	%
Random	110	15.2%
Matched	235	32.4%
"Repeated Measures"	18	2.5%
Uncontrolled	362	49.9%
	725	100.0%

Table 2.9

Joint Distribution of Smaller and Larger Class-sizes in the Comparisons Δ_{S-L}

	Larger Class-size								
Smaller Class-size	1	2	3	4-5	6-10	11-16	17-23	24-34	≥35
1	-	1	6	1	3	7	1	34	0
2		-	0	1	0	0	1	0	0
3			-	0	0	0	0	6	0
4- 5				-	0	0	1	2	0
6-10					-	8	0	5	2
11-16						-	19	44	27
17-23							-	78	106
24-34								-	197
≥35									-

The simple statistical properties of the Δ's were interesting in themselves, even though their full import required more sophisticated analysis:

Properties of Distrubution of Δ_{S-L}

a) $N = 725$.

b) Mean = .088; Median = .050

c) 40% of the Δ_{S-L} were negative; 60%, positive.

d) Standard deviation = 0.401.

e) Range: -1.98 to 2.54.

On the average, the 725 Δ_{S-L}'s were positive, i.e., over all comparisons available -- regardless of the class-sizes compared -- the results favored the smaller class by about a tenth of a standard deviation in achievement. This finding is not too interesting, however, since it is an average across many different sizes of classes compared. However, only 60 percent of the Δ's were positive, i.e., favored the smaller class in achievement. This is so, even though every effort was made to find studies spanning the full range of class-sizes from individual tutorials to huge lectures. One suspects that the odds of observing a positive Δ_{S-L} in the class-size range so often studied (15 to 40, say) were even smaller, perhaps as low as 55 percent to 45 percent.

In these rough summaries, one of the fundamental problems is revealed that has made the class-size literature so difficult for reviewers. If the relationship one seeks has only 55 to 45 odds of appearing and one looks for it without all the tools of statistical analyses that can be mustered, the chances of finding it are slight. One need not wonder why narrative reviews of a dozen or two studies produced little but confusion.

To make sense of the class-size and achievement relationship one must account for the magnitude of the Δ's and their variance in terms of the sizes of the smaller and larger classes. What was needed was a continuous quantitative model that would relate class-size $\underline{C}$ to achievement $\underline{z}$. Class-size and achievement might be expected to be related in something of an exponential or geometric fashion -- reasoning that one pupil with one teacher learns some amount, two pupils learn less,

three pupils learn still less, and so on. Furthermore, the drop in learning from one to two pupils might be expected to be larger than the drop from two to three, which in turn is probably larger than the drop from three to four, and so on. A logarithmic curve represents one such relationship:

$$z = \alpha - \beta \log_e C + \varepsilon, \tag{2.2}$$

where C denotes class-size. Since β could be zero or negative, the model in (2.2) does not preclude the data showing that class-size and achievement are unrelated or that larger classes learn more than smaller ones.

In formula (2.2), α represents the achievement for a "class" of one person, since $\log_e 1 = 0$, and β represents the speed of decrease in achievement as class-size increases. Formula (2.2) cannot be fitted to data directly because $\underline{z}$ is not measured on a common scale across studies. This problem was circumvented by calculating Δ_{S-L} for each comparison of a smaller and a larger class within a study. Then, from formulas (2.1) and (2.2) one has:

$$\begin{aligned}\Delta_{S-L} &= (\alpha - \beta\log_e S + \varepsilon_1) - (\alpha - \beta\log_e L + \varepsilon_2) \\ &= \beta(\log_e L - \log_e S) + \varepsilon_1 - \varepsilon_2 \\ &= \beta\log_e(L/S) + \varepsilon'\end{aligned} \qquad (2.3)$$

The model in formula (2.3) was particularly simple and straightforward. The values of Δ_{S-L} were merely regressed onto the logarithm of the ratio of the larger to the smaller class-size, forcing the least-squares regression line through the origin.

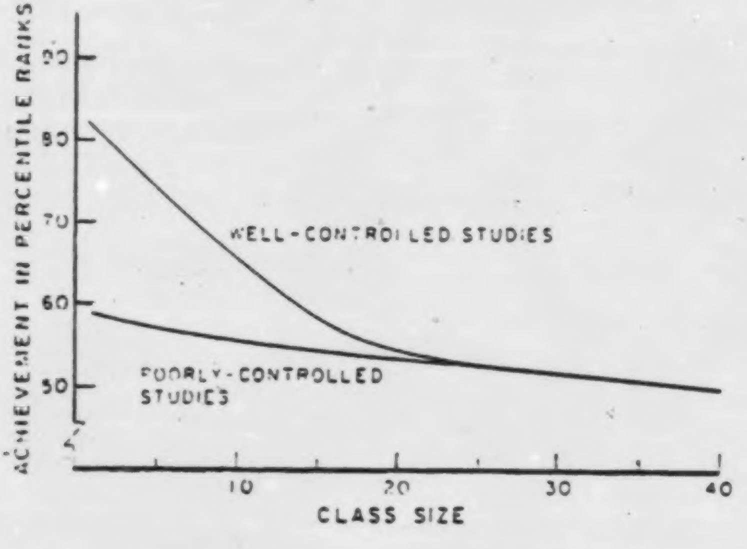


Figure (2.4) Regression lines for the regression of achievement(expressed in percentile ranks) onto class-size for studies that were well-controlled and poorly-controlled in the assignment of pupils.

The least-squares estimate of the β parameter was found to have the form:

$$\tilde{\beta} = \frac{\Sigma(\Delta_{S-L})(\log_e L/S)}{\Sigma(\log_e L/S)^2} .$$

The model in formula (2.2) was fitted to the data base as a whole and to many subdivisions of it. The strength of the relationship between class-size and achievement did not vary with characteristics of the studies (e.g., age of pupils, ability, subject taught) with one exception. The relationship was much stronger for studies in which pupils were randomly assigned to the classes of different sizes than for studies that used matched or uncontrolled assignment; thus, better controlled studies gave more positive results. Hence, we restricted our estimation of the relationship to the 100 or so Δ's that arose from the well-controlled experiments. After fitting the model in formula (2.2) to the data, estimating β and transforming $\underline{z}$ to a percentile scale, the relationships in Figure 2.4 emerged. Assuming arbitrarily that the average pupil in a class of 40 scores at the 50th percentile in achievement, his improvement in achievement as class-size is reduced as indicated by the upper curve in the figure. When he is taught in a class of 15 his achievement rises to the 60th percentile; in a group of 10, he will score at the 65th percentile; and taught by himself (class-size equal 1), he is expected to score above the 80th percentile. We concluded our report with these words:

> A clear and strong relationship between class-size and achievement has emerged. The relationship is seen most clearly in well-controlled studies in which pupils were randomly assigned to classes of different sizes. Taking all findings of this meta-analysis into account, it is safe to say that between class-sizes of 40 pupils and one pupil lie more than 30 percentile ranks of achievement. The difference in achievement resulting

> from instruction in groups of 20 pupils and groups of 10 can be larger than 10 percentile ranks in the central regions of the distribution. There is little doubt that, other things equal, more is learned in smaller classes.
>
> (Glass and Smith, 1979, p. 15)

The impact of our findings was immediate. At first the word of the findings spread informally, through face-to-face contact. A friend mentioned the study during the interview on an entirely different subject with the foreign education writer for the "London Times." An article followed, then several others as one thing led to another. The process that ensued at that point more resembled Brownian movement than linear, heirarchial dissemination. In a span of a year, synopses of the findings appeared in magazines ("Today's Education," "Psychology Today," "Forum"), newspapers ("New York Times," "Denver Post," "London Times," AP wire service), and were discussed in radio and television interviews that must have reached millions of people. The phone bagan to ring with questions and requests for documents. Parents, teachers, administrators, politicians (Pennsylvania, Georgia, Nevada, Colorado, North Carolina and Minnesota) -- they either read about the study in the popular press or heard of it from an acquaintance. Teachers unions waved the report under the noses of boards and administrators; the latter criticized it as inaccurate or hired critics to discredit it.

Sex Bias in Counseling and Psychotherapy

Smith (1980) found 34 studies of possible bias of counselors and psychotherapists toward male vs. female clients. A typical study examined experimentally the possibility that counselors and therapists varied their diagnoses, recommendations and attitudes toward their client depending on the client's gender. The 34 studies contained 60 assessments of possible sex bias.

There was wide variation in the designs used, their adequacy, and the extent to which "client" individual differences were considered. However, each study was used in the meta-analysis regardless of its qualities. Thus, the author's theoretical and methodological biases had minimal influence. The studies were rated for design quality so that the magnitude of sex bias produced by studies of different levels of design quality could be ascertained. A score or 3 was given for studies in which all experimental variables were controlled and the effects of client characteristics. A score of 2 was given to studies that merely had experimental variables under control. A score of 1 was assigned to studies in which experimental variables were uncontrolled or seriously confounded.

Methods for transforming the analytic results of the studies into a common metric followed Glass's (1978) specifications. Each dependent variable from the studies was converted into an "effect of sex bias" (ESB) according to the following formula: $ESB = (M_{Male} - M_{Female})/\hat{\sigma}$.

In a study of the effect of client gender on therapist judgment of client prognosis, for example, the mean for the prognosis given to females was subtracted from the mean prognosis given to males. The difference was divided by their average standard deviation. The resulting ESB is in the form of a normal unit deviate. An ESB of 1 indicates that the mean of the males on the dependent variable is of the magnitude of 1 SD higher than the mean of the females on that variable. In the above example, an ESB of 1 would indicate that counselors gave males a much more favorable prognosis than that given to females; in fact, the average male prognosis is more favorable than the prognosis for 84 percent of the females, assuming a normal distribution of the sex bias variable.

The ESB is standardized so that different measures can be viewed on a common, convenient metric and combined with others to form an overall picture of the sex bias effect. The dependent measures were arranged so that a positive ESB always meant bias against females or against nontraditional, noncomformist, or androgynous actions, decisions, or labels. A negative ESB indicated bias in favor of females or nonconforming, nonstereotypic goals. One study illustrates this process. Price and Borgers (1977) compared counselors' ratings of appropriateness of course selection for boys and girls. The mean appropriateness rating given to boys was 3.5. The mean appropriateness rating given to girls was 3.45. The average standard deviation for boys and girls was .95. The ESB was .05. That is, the rated appropriateness was biased against females by a magnitude of .05 *SD* units, a very small amount.

Transformation of dependent measures into ESBs was straightforward when means and standard deviations were given. When *t*, *F*, or chi-square statistics were given, estimates of $\hat{\sigma}$ were found by backward solution of statistical formulas. For example, an estimate of $\hat{\sigma}$ can be found from a study in which only a value for *F*, the *n*'s, and the treatement means are reported by using the following steps:

$$MS_{\text{Between}} = \Sigma n_j (M_j - M_{..})^2 / (J - 1)$$
$$MS_{\text{Within}} = \frac{MS_{\text{Between}}}{F}$$
$$\hat{\sigma} = \sqrt{MS_{\text{Within}}}$$

where $\underline{J}$ is the number of groups and n_j is the number of cases per group. More complicated procedures permitted the estimation of $\hat{\sigma}$ from designs with blocking variables and covariates, as specified by Glass (1978) and elaborated in McGaw and Glass (1980). Special problems arose in the calculations of ESB when the researcher reported only significance levels of effects; when, for example, the researcher stated that client sex produced no significant differences on the dependent variable. In this case, an ESB of zero was entered for that variable.*

Another problem was encountered in studies that reported item-by-item significance tests on sex-role stereotyping measures. The item-level data were converted to ESBs, and the average ($\overline{\text{ESB}}$) for the item set was recorded for that study. Except for these few studies in which multiple item-level data were averaged, the practice was to record an ESB for each dependent measure that the researcher reported.** Table 2.10 contains the ESBs calculated for the studies.

The ESB measures were accumulated by the domain under investigation (counseling or psychotherapy) and by the construct measured (attitudes, judgments, or behaviors) and for other variables of interest.

---

*A check on this procedure was conducted after the meta-analysis was completed. Neither altering the procedure nor eliminating these findings from the summary changed the final $\overline{\text{ESB}}$ by more than a fraction.

**A later check on the effect of $\overline{\text{ESB}}$ calculated at the level of the dependent measure and at the level of the study showed no differences in the magnitude of effect.

The resulting summary statistics are contained in Table 2.11.The means, standard deviations, and the number of effects are presented, along with the standard error of effects $\left(\frac{\sigma_{ESB}}{\sqrt{n}}\right)$. Whether the number of studies in a meta-analysis should be considered the entire population of studies on a topic or rather a sample of a hypothetical population of such studies is problematic. If the latter is true, then inferential statistics might be appropriately applied to the effect-size measures. However, appropriate sampling distributions for inferential statistics in meta-analysis have yet to be evaluated. Presentation of the standard error of effects allows the reader a rough-and-ready measure of the significance of difference of the means of two contrasting conditions (e.g., $\overline{ESB}$ for well-controlled studies vs. $\overline{ESB}$ for poorly controlled studies). A difference in means less than two standard errors in magnitude was deemed unreliable and did not figure into the discussion of results.

Table 2.11 contains the summary statistics for the sex-bias meta-analysis. The overall mean of ESBs is given along with the mean for each construct, domain, the source of the study, and the validity of the design.

The results are clear. There is no evidence for the existence of counselor sex bias when the research results are taken as a whole. The average ESB is -.04, indicating that the counselor bias is near zero or even slightly in favor of women and nonstereotyped actions for women. The size of the sex-bias effect does not change from construct to construct. Attitudes, judgments, and behaviors all show about the same size of effect. Considered separately, the findings labeled clinical stereotypes produced an $\overline{ESB}$ of .24, which recapitulates the conventional wisdom that clinicians hold negative stereotypes about women. When the standard error of effects is used to evaluate this, one finds that the $\overline{\text{ESB}}$ for stereotypes is not reliably different from the $\overline{\text{ESB}}$ of the data as a whole.

The analysis of sex bias found in journals as opposed to dissertations is extremely interesting. Journal articles were much more likely to show bias against women. Dissertations showed the opposite. One is tempted to suppose that dissertations are more poorly designed and executed and therefore less likely to be published. That supposition would be incorrect, as the average rating of design quality was slightly higher for dissertations than for journals (2.57 and 2.16, respectively). The best designed studies -- those in which experimental variables were well controlled and provision was made to isolate gender effects from personal characteristics -- yielded results opposite to those of the sex-bias hypothesis. Studies with moderate validity -- controlled variables but no provision for gender and case distinctions — averaged zero on the ESB variable. Studies with poor controls or severe confounding of variables yield the results most supportive of the sex-bias hypothesis.

Analysis of interactions of variables failed to yield reliable results, with one exception. There was a statistically significant interaction between design quality and publication status, but not in the predictable direction. Table 2.12 contains the $\overline{\text{ESB}}$ and standard error of $\overline{\text{ESB}}$ for the Design Quality X Source of Publication interaction. Studies published in journals were more likely to show the effect of sex bias, regardless of the quality of their research design. Viewed another way, studies most likely to be submitted or accepted for publication tended to be those that demonstrated the sex-bias effect, their design quality notwithstanding.

Table 2.10

Author, Source, Domain, Construct, Type of Effect, and Effect of Sex Bias (ESB) of Studies

Author	Source (*Dissertation* or *Journal*)	Domain (*Psychotherapy* or *Counseling*)	Construct (*Attitude*, *Judgment*, or *Behavior*)	Validity	Type of effect	ESB
Aslin (1975)	D	P	A	2	Sex stereotypes	.23
Broverman, Broverman, Clarkson, Rosenkrantz, & Vogel (1970)	J	P	A	2	Sex stereotypes of mentally healthy persons	.56
Friedersdorf (1970)	D	C	A	2	Sex stereotyped interests	.00
Hayes & Wolleat (1978)	J	C	A	2	Sex stereotypes	−.31
Maslin & Davis (1975)	J	C	A	2	Sex stereotypes	.56
Lesser (1975)	D	C	A	2	Acceptance of self-orientation	−1.03
Maxfield (1976)	D	P	A	3	Sex stereotypes	.00
Neulinger (1968)	J	P	A	2	Sex stereotypes	.60
Smith (1973)	D	C	A	3	Sex stereotypes	.01
Wirt (1975)	D	C	A	3	"Evaluation"	−.68
					"Potency"	.63
					"Activity"	−.87
Abramowitz, Abramowitz, Jackson, & Gomes (1973)	J	C	J	3	Psychological adjustment	.14
Abramowitz et al. (1976)	J	P	J	2	Prognosis	−.22
Abramowitz et al. (1975)	J	C	J	2	Psychological adjustment	.01
Billingsly (1977)	J	P	J	3	Treatment goals	.00
Borgers, Hendrix, & Price (1977)	J	C	J	2	Appropriateness of vocational choice	−.10
Coen (1975)	D	P	J	3	Desire to treat	−.46
					Degree of impairment	.06
					Prognosis	−.30
Donahue (1976)	J	C	J	1	Remuneration of vocational choice	.58
					Education required for vocational choice	.21
					Supervision required for vocational choice	.61

Table 2.10 (countinued)

Author	Source (*Dissertation* or *Journal*)	Domain (*Psychotherapy* or *Counseling*)	Construct (*Attitude*, *Judgment*, or *Behavior*)	Validity	Type of effect	ESB
Freedman (1976)	D	P	J	2	Personality type	.15
					Degree of disturbance	−.26
					Treatment type	−1.15
					Readiness for therapy	−.30
					Willingness to treat	−.42
Goldberg (1976)	D	P	J	3	Conventionality of chosen occupation	.00
Hill, Tanney, Leonard, & Reiss (1977)	J	C	J	3	Seriousness of problem	.22
					Ability to profit from counseling	−.17
					Attractiveness as client	.04
					No. sessions needed	.00
Lesser (1975)	D	C	J	2	Acceptance of self-orientation	−.03
Lewittes, Moselle, & Simmons (1973)	J	P	J	2	Degree of pathology	.00
Maxfield (1976)	D	P	J	3	Degree of disability	.00
					Recommendation for counseling	−.61
					Prognosis	−.54
					Recommendation for hospitalization	.00
					Diagnosis	.00
Price & Borgers (1977)	J	C	J	2	Appropriateness of course choice	.05
Smith (1973)	D	C	J	3	Need for further counseling	.24
Smith (1974)	J	C	J	3	Prediction of academic success	−.04
					Recommended occupation	.05
Thomas & Stewart (1971)	J	C	J	2	Acceptance	.08
					Appropriateness of career goal	.66
					Need for further counseling	.32
Hill (1975)	J	C	B	2	Counselor behaviors (empathy, etc., combined)	.11
Hill, Tanney, Leonard, & Reiss (1977)	J	C	B	3	Empathy	−.23
Libbey (1976)	D	P	B	3	Positive emotion	−1.13
					Specificity	−.52
					Confrontation	−.02
Petro & Hansen (1977)	J	C	B	2	Affective sensitivity	.29
Schlossberg & Pietrofesa (1973)	J	C	B	1	Sex-biased statements	1.68
Stengel (1976)	D	C	B	2	Empathy	.00
					Warmth	.00
					Genuineness	.00
Wirt (1975)	D	C	B	3	Empathy	.00
					Positive regard	−.21
					Genuineness	−.87

Table 2.11

Summary Data From Sex Bias Meta-Analysis—Means, Standard Deviations, Number of Effects of Sex Bias (ESB), and Standard Error of $\overline{ESB}$ ($\hat{\sigma}_{\overline{ESB}}$)

Variable	$\overline{ESB}$	σ_{ESB}	n_{ESB}	$\hat{\sigma}_{\overline{ESB}}$
Construct				
Attitudes	−.03	.59	12	.17
Judgments	−.03	.35	35	.06
Behaviors	−.07	.66	13	.18
Domain				
Psychotherapy	−.18	.43	24	.09
Counseling	.05	.48	36	.08
Source				
Journals	.22	.41	28	.08
Dissertations	−.24	.46	32	.08
Design validity				
High	−.18	.38	30	.07
Medium	−.01	.43	26	.08
Low	.77	.63	4	.32
Total	−.04	.47	60	

Table 2.12

Interactions of Design Validity and Source of Study—Means, ns, and Standard Errors of Sex Bias

Validity of design	Source of study		
		Journals	Dissertations
Low	$\overline{ESB}$ =	.77	No cases
	$\hat{\sigma}_{\overline{ESB}}$ =	.32	
	n_{ESB} =	4	
Medium	$\overline{ESB}$ =	.19	−.23
	$\hat{\sigma}_{\overline{ESB}}$ =	.09	.13
	n_{ESB} =	14	12
High	$\overline{ESB}$ =	.00	−.25
	$\hat{\sigma}_{\overline{ESB}}$ =	.05	.09
	n_{ESB} =	9	21

Drug Therapy for Psychological Disorders

Miller (1978; also see Smith, Glass and Miller, 1980) sought to integrate a fragmented and widely scattered empirical literature on the effects of drug therapy on persons with debilitating psychological disorders. A conventional wisdom had long pervaded the field; and it both reflected and supported the political equilibrium that psychiatrists and psychologists had struck. Ask most mental health practitioners and they would have told you that verbal psychotherapy practiced by itself on the seriously disturbed (schizophrenic, psychotic) is a waste of time; but combine it with drug treatment (which is effective in isolation) and the synergistic combination is much more beneficial than the sum of their separate contributions. Psychologists who believed this would serve at the pleasure of psychiatrists, who are empowered by law to prescribe pharmaceuticals.

Miller found several thousand experimental studies that bore on the question of the relative efficacy of drug and psychotherapy effects. Most of these were clinical trials comparing drugs against placebos. From this huge literature, Miller samples at random about fifty studies. The remainder of the literature comprised about 125 experiments that compared drugs and psychotherapy in various odd combinations (e.g., drug-plus-psychotherapy vs. drug vs. psychotherapy; drug-plus-psychotherapy vs. placebo).

Miller calculated the standardized average difference on the dependent variable for each of the outcomes measured in the experimental comparisons in the   studies. Nearly 550 effects were thus calculated. Summaries of the averages appear in Table 2.13. There one sees, for example, that in 55 comparisons of verbal psychotherapy with an untreated control group or placebo, the psychotherapy group averaged .30 standard deviation units higher on the outcome measure. In 94 comparisons of drug-plus-psychotherapy with psychotherapy alone, the former averaged .44 standard deviation units higher than the latter on the dependent variables measured in the experiments. Table 2.13 gives a parametric structure for the comparisons with numeric parameters to be estimated from the data. Such quantification is required of what are essentially quantitative questions about separate and interactive effects of drugs and psychotherapy. Narrative and box-score summaries are quite at a loss to cope with such problems.

Consider now the problem of combining data in Table 2.13 to obtain estimates of the parameters. That the drug-plus-psychotherapy vs. drug comparison, which estimates $\Psi + \eta$, is a full one-tenth standard deviation larger than the .30 estimate of Ψ from the first line of the table might lead one to believe that η is positive; but the comparison of the estimates of $\delta + \eta$ and δ (being .44 and .51, respectively) reverses this impression. Parameter estimation by inspection in this way is too arbitrary and confusing. Several comparisons in the table contain information about the same parameters; it seems reasonable that every source of information about a parameter should be used in estimating it. A complete and standard method of combining the data in Table 2.13 into estimates of the parameters is needed. Such a method is suggested when one recognizes that the two middle columns of Table 2.13

Table 2.13

Average Effect Sizes from Various Experimental Comparisons

Made in the Experiments on Drug and Psychotherapy

Comparison	Parameter(s) Estimated	Average ES	No. of ES's
Psychotherapy vs. No-Treatment or Placebo	ψ	.30	55
Drug Therapy vs. No-Treatment or Placebo	δ	.51	351
Drug & Psychotherapy vs. Drug	$\psi + \eta$	.41	10
Drug & Psychotherapy vs. Psychotherapy	$\delta + \eta$	.44	94
Drug vs. Psychotherapy	$\delta - \psi$	.10	7
Drug & Psychotherapy vs. No-Treatment or Placebo	$\delta + \psi + \eta$	.65	49

Note. ψ denotes the separate or "main" effect of psychotherapy; δ denotes the separate effect of drug therapy; and η denotes their interaction.

constitute a system of linear equations, three of them independent and containing three unknows (ψ δ and η). The method of least squares statistical estimation can be applied to obtain estimates of the separate and interactive effects of drug and psychotherapy. The estimates obtained by application of least-squares methodology to the data in Table 2.13 are as follows:

ψ , the separate effect of psychotherapy = .31

δ , the separate effect of drug therapy = .42

η , the interactive effect of drug-plus-psychotherapy = .02

Each effect is expressed on a scale of standard deviation units. Thus, the data of Table 2.13 lead to the conclusion that with the groups of clients studied psychotherapy produces outcomes that are about one-third standard deviation superior to the outcomes from placebo or untreated control groups. The drug effect is only about a third greater than the psychotherapy effect. An effect of $.31s_x$ will move an average client from the middle of the control group distribution to about the 62nd percentile; an effect of .42 would move the average client to only about the 66th percentile. The effects of the two therapies were conducted for only half the time it took to conduct the psychotherapies (2.6 months vs. 6.1 months). Any careful assessment of the relative value of drug and psychotherapy will take both effects and costs into account.

Arguments over the relative value of drug and psychotherapy will be simpler for the fact that the interactive effect of combining the two therapies is virtually zero (η = .02). This must not be misunderstood as implying that drug-plus-psychotherapy is ineffective;

far from it. The near zero interaction effect means that when drug and psychotherapy are combined, one can expect benefits equal to the sum of the separate drug and psychotherapy effects (.31 + .42 = .73), not more or less.

CHAPTER THREE

FINDING STUDIES

Reviewing and integrating a research literature begins, obviously enough, with the literature — often a widely-scattered, variegated landscape of articles, theses, project reports and whatever. Jackson (1978) showed how this first step was occasionally taken rather uncertainly by reviewers. Of 36 reviews that Jackson analyzed, only one reported having searched the literature with the help of indexes like Psychological Abstracts or Dissertation Abstracts; only three of the 36 reviews reported searching bibliographies of previous reviews of the topic. Whether reviewers do not take such obvious steps in finding studies or take them but neglect to say so may be immaterial from the reader's point of view; in either case it is difficult to judge whether the studies being reviewed represent most of the existing evidence on the question or only an unrepresentative portion. Earlier we likened meta-analysis to survey research; thus, finding studies is comparable in importance to sampling frames and methods in survey design and analysis. Locating studies is the stage at which the most serious form of bias enters a meta-analysis, since it is a potential bias whose impact is difficult to assess. The best protection against inestimable sources of bias is a thorough description of the procedures used to locate the studies that were found so that the reader can make an intelligent assessment of the representativeness and completeness of the data base for a meta-analysis.

As an example of the lengths to which one might sometimes have to go to feel confidence of having done a thorough job of finding relevant studies, consider Miller's (1978) experiences in reviewing an enormous literature on the psychological effects of drug therapy.

> "To draw conclusions about the entire realm of clinical drug research on psychological disorders, a sample was taken from the large number of existing drug therapy studies. An attempt was made to draw a representative sample of all published clinical drug trials on mentally ill humans reported in the English language literature between 1954 and 1977.
>
> The only design requirement for inclusion in the sample was that studies employ a no-drug treatment or a placebo control group. Though previous reviewers were admonished for inclusion requirements that were, in this author's opinion, too restrictive (e.g., including only double-blind placebo controlled studies), this somewhat arbitrary line was drawn because of a conviction that without a control group, spontaneous symptom remission rampant in psychiatry would be recorded as a drug effect. Case studies, experiential reports, pre-post designs, and drug versus drug studies were therefore omitted.
>
> To identify more clearly the domain from which to sample, further restrictions were imposed on selection of potential studies. Studies of patients whose primary diagnosis was somatic were excluded. Thus omitted were studies of drugs used to treat patients for organic brain syndrome, epilepsy, phenylketonuria, minimal brain damage, or Down's Syndrome, and studies of patients with psychophysiological disorders (asthma, backache, acne, ulcer, enuresis, angina, etc.). This criterion did not exclude studies whose primary focus was examination of neurotic or psychotic patients or patients with character disorders whose somatization of symptoms led to physiological illness.
>
> All studies of normal subjects and all studies that used only physiological outcomes (e.g., blood plasma levels of amines, EEG's, urinalysis) were omitted. Lastly, studies of toxic psychosis (e.g., drug induced psychosis) or model psychosis (e.g., using hallucinogens) were not examined.
>
> A Medical Literature and Retrieval System (MEDLARS) search from the University of Colorado Medical Center computer search facility generated all research meeting specified criteria catalogued between January 7, 1966 and January 30, 1977. (The search specifications appear in Table 3.1.) The facility catalogues all studies from approximately 2,400 journals.

Studies could not be suppressed by design characteristics or outcome variables so though all listed studies met the inclusion requirements there was an unspecified number of studies listed that met the exculsion requirements as well (e.g., there were some uncontrolled studies and studies designed to assess only bio-chemical outcomes of drug administration). Approximately 1,100 studies were located by the MEDLARS search.

Several studies were selected at random from the MEDLARS print-outs. As the referenced articles were located and read, it became clear that many studies lacked control groups. Titles containing no allusion to the existence of a control group (via such key words as "double-blind," "crossover," "controlled," or "placebo") portended studies lacking this crucial ingredient. Therefore, to reduce reference retrieval time by directing gathering efforts toward studies very likely to have control groups, articles with titles containing the above-mentioned key words became the primary focus of the random sample. Forty such studies were randomly chosen from the MEDLARS bibliography.

From the psychopharmacological literature prior to January 1, 1966, the period not covered by MEDLARS, a random sample of about fifty studies was taken from bibliographies of comprehensive review articles on the efficacy of drug treatment in psychiatric cases and from studies listed in Psychological Abstracts between 1954 and 1966 under the heading Therapy/Drugs. These review articles and the number of bibliographical references made in each are presented in Table 3.2. Shown as the last reference in Table 3.2 is the number of studies sampled from the 1954-1966 Psychological Abstracts that became part of the pool of pre-1966 references from which studies were sampled.

Once again the emphasis on title terminology that was likely to indicate the use of a control group was applied to selection of studies from these bibliographies.

The selection of the ninety or so articles (fifty articles from the 1954 to 1966 literature; forty articles from the 1967 to 1977 literature) was stratified so that approximately equal numbers would be represented in three major drug categories: antipsychotic, antianxiety, and antidepressant. Once these articles were assembled a few articles were added to assure that major well-known studies and very recently published articles (February and March, 1977) were not overlooked. Ninety-six articles or books studying the effects of drug therapy were thus collected, read and coded."

(Miller, 1978, pp. 31-36.)

Table 3.1

SEARCH FORMULATION - PSYCHOPHARMACOLOGIC RESEARCH IN HUMANS: CLINICAL TRIALS

1) Affective Disturbances F3.126.56
   - Anxiety
   - Depersonalization
   - Depression

2) Aggression
   - Torture
   - Violence

3) Neuroses
   - Depression
      - Depression, Reactive
         - Depressive Neuroses
      - Hypochondriacs
         - Hypochondriacal Neuroses
         - Munchausen Syndrome
      - Hysteria
         - Conversion Reaction
            - Globus Hysterious
         - Dissociative Reaction
            - Duel Personality
               - Multiple Personalities
            - Hysterical Neuroses
      - Neuroesthenia
      - Neurocirculatory Asthenia
      - Neuroses Anxiety
         - Anxiety, Castration
         - Anxiety, Separation
            - Homesickness
      - Neuroses, Obsessive-Compulsive
         - Kleptomania
         - Obsession
         - Trichotillomania
      - Neuroses, Post-Traumatic
      - Neuroses, War
      - Phobias
         - Agraphobia
         - Claustrophobia
         - Phobia Neuroses

4) Compulsive Behavior F3.126.208
   - Firesetting Behavior
   - Gambling
      - Risk Taking

   - Obsessive Behavior
   - Smoking

5) Alcholism C21.613.53.270
   - Psychoses, Alcoholic
      - Delirium Tremens
      - Kersakoff's Syndrome

6) Psychoses F3.709.755
   - Depression, Reactive, Psychotic
   - Folie a Deux
   - Paranois
   - Psychoses, Involutional
      - Involutional Paranoid State
      - Melancholic, Involutional
      - Paraphrenia
   - Psychoses, Manic-Depressive
   - Schizophrenia
      - Schizophrenia, Catatonic
      - Schizophrenia, Childhood
         - Autism, Early Infant
      - Schizophrenia, Hebephrenia
      - Schizophrenia, Latent
      - Schizophrenia, Paranoid

7) Personality Disorders
   - Alcoholism
      - Skid Row Alcoholic
   - Antisocial Personality
      - Sociosyntonic Personality Disorder
   - As If Personality
   - Cyclothymic Personality
   - Hysterical Personality
   - Inadequate Personality
   - Obsessive-Compulsive Personality
   - Paranoia
      - Paranoic Personality
   - Passive-Aggressive Personality
      - Passive-Dependent Personality
   - Schizoid Personality
   - Tension-Discharge Disorders
      - Impulse-Ridden Personality

8) 1 or 2 or 3 or 4 or 5 or 6 or 7

9) 8 and Drug Therapy

10) Clinical Research or Research Design

11) 9 and 10

12) 11 and Placebos

13) 11 and (control or controlled)

14) 12 or 13

Table 3.2

Bibliographic References from Which Most Pre-1966 Studies were Sampled

Reference	Topic	Number of Studies Listed in Bibliography
Azcarate (75)	Anti-Agression	43
Davis (65)	Anti-Depression	410
Davis et al. (68)	Anti-Depression	369
Hollister, L. (69)	All Drugs	120
Hollister, L. (73)	All Drugs	241
Itil, T. (75)	Anti-Aggression	81
Klein and Davis (69)	a) Mood Stabilizer	472
	b) Minor Tranquilizer	185
	c) Anti-Psychotic	420
Klerman and Cole (65)	Anti-Depression	341
Morris and Beck (74)	Anti-Depression	185
Sheard, M. (75)	Anti-Agression	60
Psychological Abstracts (1954-1966)	Therapy/Drugs	26
	TOTAL	2,963

Miller's example has been reported here in rather more detail than may seem polite to the reader to make a point. Documenting the methods used in finding research literature takes more space than custom traditionally allocates to describing one's search. How one searches determines what one finds; and what one finds is the basis of the conclusions of one's integration of studies. Searches should be more carefully done and documented than is customary.

The Landscape of Literature

Scholarly, empirical literature in the social sciences and applied fields can be found in either primary or secondary sources. By primary sources is meant the archival periodical literature — "the journals," hundreds, perhaps thousands, of them from all over the world. Dissertations and theses are also regarded as primary sources, as well as "fugitive" literatures of government reports, papers from scholarly meetings, reports to foundations, public agencies and the like.

Secondary sources cite, review and organize the material of the primary sources; they include review periodicals (e.g., Psychological Bulletin, Review of Educational Research, Sociological Review), periodocal reviews (Encyclopedia of the Social Sciences, Encyclopedia of Educational Research), and various abstract and citation archives.

Abstracts in Anthropology,
Child Development Abstracts and Bibliography
Current Index to Journals in Education
Dissertation Abstracts International
Education Index
Government Reports Announcements & Index
Index Medicus
Index of Economic Articles

Interagency Panel Information System
International Bibliography of Economics
International Bibliography of Political Science
International Political Science Abstracts
Journal of Economic Literature
Library of Congress Catalog
National Clearinghouse for Mental Health Information
National Institute for Mental Health Grants and Contracts
  Information System
National Technical Information Service
Psychological Abstracts
Research in Education
Smithsonian Science Information Exchange
Sociological Abstracts

Some systems are computerized and quite sophisticated. For example, the Educational Resources Information Center operated by the National Institute of Education is a remarkable service that not only indexes and abstracts the published literature in education (see Current Index to Journals in Education) but the fugitive literature as well (see Resources in Education). More significantly, ERIC is a system organized around a thesaurus of topic descriptors assigned by experienced staffs of readers of the documents; this feature represents a significant advance over indexes that depend on author selected descriptors or the key words of titles.

Perhaps we have said enough at this level. The reader who has gotten this far is unlikely to be a stranger to modern libraries and the delights that they hold. And the technology of information storage and retrieval is advancing so rapidly that whatever detail we might give here is likely soon to be out of date.

Literature Searches in Meta-analyses

Our topic is the methodology of meta-analysis, so in the remainder of this chapter we shall limit ourselves to a couple of considerations about literature searching that bear directly on meta-analysis.

Reliability of Literature Searches

No matter how ambitious and sophisticated are one's efforts to find all empirical research on a topic, the aspiration to find everything must be inevitably frustrated. There is simply too much literature in too many strange places to find it all. But reviewers can do a better job than they typically have done. The arbitrary exclusion of vast amounts of literature (e.g., excluding all dissertations or all fugitive manuscripts in ERIC) is unsound and bespeaks more faintness of heart than intelligence of judgment. Nevertheless, the most conscientious efforts fall short of perfect. There is less reliability in searching for research studies than would be tolerable in survey research, for example; but it is an especially intransigent sort of unreliability for which we have no facile answers.

We tested the reliability of four large study indexes by computerized search on descriptors for "group homes for delinquents." The four indexes were ERIC (Educational Resources Information Center), Psychological Abstracts, Dissertation Abstracts, and Council for Exceptional Children Abstracts. A total of 27 different studies were found. But they were distributed according to the following cross-classification.

Numbers of Listings from
Different Data Bases

Search on: (Achievement (w) Place) and (Teaching (w) Family)
(group homes for delinquents)

	ERIC	PSYCHOLOGICAL ABSTRACTS	DISSERTATION ABSTRACTS	CEC ABSTRACTS
ERIC	8	2	—	3
PSYCHOLOGICAL ABSTRACTS	2	22	2	9
DISSERTATIONS ABSTRACTS	—	2	4	—
CEC ABSTRACTS	3	9	—	18
UNIQUE	5	11	2	9

For example, of 8 studies on the topic found in the ERIC system, two were also listed in Psychological Abstracts, and three also appeared. in the CEC Abstracts. Five of the 8 ERIC studies did not appear in any of the other three indexes. The greatest proportion of redundancy appears to be between Psychological Abstracts and CEC Abstracts on this topic. The above table gives one pause. Perhaps the social and behavioral sciences need indexes of indexes!

Publication Bias and Meta-analysis

Meta-analyses may be thought of as a type of survey research. The goal of the meta-analyst should be to provide an accurate, impartial, quantitative description of the findings in a population of studies on a particular topic. This may be done by exhausting the population or sampling representatively from it. No survey would be considered valid if a sizable subset (or stratum) of the population was not represented in the cumulative results. Neither should a meta-analysis be considered complete if a subset of its population is omitted. One very important subset of evidence is the subset of unpublished studies. To omit dissertations and fugitive research is to assume that the direction and magnitude of effect is the same in published and unpublished works.

The most radical criticism of the assumption of equivalence is the old saw that the published literature only represents the five percent of false positives in a population of studies wherein the null hypothesis is true. That is, the published stratum and the unpublished stratum have opposite average effects, and a meta-analysis containing only published studies would be wholely unrepresentative of the population. Rosenthal (1979) effectively countered this attack by mathematical demonstration of the numbers of studies which would have been languishing in file drawers to make up the 95 percent null results. The existence of such huge numbers is considered implausible.

The results of meta-analyses which did represent both published and unpublished literature provide further evidence on the assumption of equivalence. Table 3.3 contains the results of 12 such meta-analyses. In every one of the ten instances in which the comparison can be made, the average experimental effects from studies published in journals is larger than the corresponding effect estimated from theses and dissertations. That is, if one integrates only "published" (meaning journal published) studies, the impression of support for the favored hypothesis is artificially enhanced over what would be seen if the entire

Table 3.3

Relationship Between Source of Publication and Findings in 12 Meta-Analyses of Experimental Literatures

			Source of Publication			
nvestigator(s)	Topic		Journal	Book	Thesis	Unpubl.
avale ('79)	Psycholinguistic training	n:	13		16	5
		$\overline{\text{ES}}$.:	.50		.30	.37
artley ('77)	Computer-based instruc.	n:	34		13	34
		$\overline{\text{ES}}$.:	.36		.28	.54
	Tutoring	n:	9		47	17
		$\overline{\text{ES}}$.:	.77		.40	1.05
osenthal ('76)	Experimenter bias	n:	25		50	
		$\overline{\text{ES}}$.:	1.02		.74	
nith ('80a)	Sex bias in psychotherapy	n:	28		32	
		$\overline{\text{ES}}$.:	.22		-.24	
nith ('80b)	Effects of aesthetics educ. on basic skills	n:	29		164	56
		$\overline{\text{ES}}$.:	1.08		.48	.50
rlberg ('79)	Spec. ed. room placement vs. reg. room placement	n:	146	17	45	114
		$\overline{\text{ES}}$.:	-.09	-.01	-.16	-.14
	Resource room plac. vs. reg. room place.	n:	33	6		
		$\overline{\text{ES}}$.:	.32	-.09		

Table 3.3 (continuned)

ller ('79)	Drug therapy of psych. disorders	n:	336	21		
		$\overline{\text{ES}}$.:	.49	.56		
arold ('79)	Effects of TV on anti-social behav.	n:	262	120	96	13
		$\overline{\text{ES}}$.:	.40	.14	.18	.23
SUBTOTALS		n:	1025	177	473	268
		$\overline{\text{ES}}$.:	.38	.18	.30	.27
mith, Glass & Miller ('80)	Psychotherapy	n:	1179	42	483	61
		$\overline{\text{ES}}$.:	.87	.80	.66	1.96
TOTALS		n:	2204	219	956	329
		$\overline{\text{ES}}$.:	.64	.30	.48	.58

literature were integrated (i.e., journals, books and dissertations). The bias in the journal literature relative to the bias in the dissertation literature is not inconsiderable. The mean effect size for journals is .64 as compared with .48 for the dissertation literature; hence, the bias is of the order of [(.64 - .48)/.48] 100% = 33%. Thus, findings reported in journals are, on the average, one-third standard deviation more favorably disposed toward the favored hypotheses of the investigators than findings reported in theses or dissertations.

Comparisons of average effect sizes among other sources of publication are less clear, in part perhaps, because of the amibguity in labels such as "unpublished" or "book." In four of eight instances, the average effect size for journals was larger than for unpublished studies. Unpublished studies seemed to divide along the following lines: one large group of old unpublished studies, containing unremarkable results that never caught anyone's attention, and a smaller group of new studies circulating through the "invisible college" while waiting to be published.

In the meta-analysis of sex bias in counseling and psychotherapy (Smith, 1980a), not only the magnitude but the direction of effect was different in published and unpublished studies. A positive effect size indicated the biasing effect of counselor attitudes, judgments, and behaviors against female clients or against non-stereotyped roles for females. The effect size from published studies was .22, demonstrating counselor bias against females. The effect size from unpublished studies was -.24 demonstrating counselor bias in favor of females.

From these data it is appropriate to conclude that failing to represent unpublished studies in a meta-analysis may produce misleading generalizations.

To omit dissertations because of their assumed lack of rigor is also unwarranted. Only after the studies have been quantified and their results transformed to effect size measures can it be determined whether published studies on a topic were more rigorously designed than were unpublished studies and whether rigor of design related to magnitude of effect. In the psychotherapy meta-analysis (Smith, Glass, and Miller, 1980), there was no reliable difference in the rigor of design of published versus unpublished studies. In the sex-bias meta-analysis (Smith, 1980b), the published studies that showed bias against females actually had *less rigorous* designs than did studies (either published or unpublished) which showed no bias against females.

To make these decisions *a priori* may inject arbitrariness and bias into the conclusions. If meta-analysis offers any improvement over traditional methods of reviewing research, it is precisely in the area of removing these sources of arbitrariness to arrive at an impartial and representative view of "what the research says."

CHAPTER FOUR

DESCRIBING, CLASSIFYING AND CODING RESEARCH STUDIES

Meta-analysis is the statistical analysis of research which works with research reports as its raw material. Thus, meta-analysis entails the quantitative description of the characteristics and findings of studies; this quantification usually involves measurement in its metric aspects (e.g., in what year was this study done? What is the sample size on which $\underline{r}_{xy}$ is based?) as well as its nominal or coding function (were initial differences corrected by analysis of covariance ? Yes = 1, No = 2). Since meta-analysis entails the measurement of study characteristics and findings, many concerns that apply to measurement more generally (e.g., reliability, validity) apply to measurement as applied in meta-analysis.

Consider the example in Table 4.1. There are recorded the characteristics and findings of about twenty correlational studies of the relationship between teachers' "indirect" teaching style (non-authoritarian, encouraging discussion instead of lecturing) and pupils' learning. For example, in Study #13 (Torrance and Parent, 1966), the indirectness of ten teachers' style was correlated with their pupils' mathematics achievement for a year-long course at the high-school level; the data were reported in the form of a Spearman rank-order correlation coefficient, which is itself the best estimate of the Pearson $\underline{r}$. The reported correlation of teacher indirectness and pupil achievement was +.32, pupils of more indirect teachers learned more math.

On the face of the problem, there are six variables or characteristics descriptive of each study: the number of teachers studied (the sample size, in effect), the duration of the period of instruction, the subject tested, the grade-level of the pupils, the form of the originally reported findings (_r_'s, _F_'s, etc.), and an estimate of the correlation on the Pearson _r_ scale. If one probes deeper, even more characteristics of studies are apparent or can be inferred from the research reports. For example, the year in which the study was reported appears in Table 4.1 and could be an interesting property of studies in a field subject to fads and trends. The identity of the researcher is known, and sometimes other charactersitics can be inferred from such knowledge, e.g., Has this researcher done several studies or only one? Has he taken a public position on what this research ought to show? How many of the researchers are related as mentor-to-student or colleague-to-colleague? Moreover, variables that appear simple and straightforward reveal unexpected complications after a closer look. Take, for instance, "grade level" in Table 4.1. A study in which _X_ and _Y_ are correlated for students and teachers are spread among several grades (across fourth, fifth and sixth but averaging grades five, say). It may be necessary, then, to code both the average or modal grade of pupils represented and the range of grades as separate characteristics of the studies. Measurement of study findings is likewise complex. It is necessary to transform the findings of each study to a common scale of Pearson's _r_ so that comparisons and contrasts can be made; but studies come reported in a bewildering variety of odd statistics. For example, in Study #11, Weber

Table 4.1

Results of Studies on the Relationship Between Teacher Indirectness and Pupil Achievement
(After Gage, 1976)

	Study	No. of Teachers	Duration of Teaching	Learning Tested Subject	Grade Level	Reported Statistics	Equivalent Value in Terms of r_{xy}
1.	Flanders (1970)	15	2 semesters	language skills number skills	2	r = -.073	-.073
2.	Flanders (1970)	16	2 weeks	social studies	4	r = .308	.308
3.	Flanders (1970)	30	2 semesters	Composite MAT	6	r = .224	.224
4.	Flanders (1970)	15	2 weeks	Social Studies	7	r = .481	.481
5.	Flanders (1970)	16	2 weeks	Mathematics	8	r = .428	.428
6.	Cook (1967)	8	2 semesters	Discussion - Lab Work	10		.09 .07
7.	Furst (1967)	15	4 one-hour lessons	Economics	10, 12	$F_{1,13} = 7.15$	.11 .26
8.	Medley-Mitzel (1959)	49	2 semesters	Reading	3-6	r = .20	.20
9.	Powell (1968)	9	2 semesters	Composite SRA Reading Arithmetic	3	F=5.85 F=10.68 F=1.30 df=1,164	.23 .11 .31
10.	Snider (1966)	17 (10 in analysis)	2 semesters	Science	12	Mann-Whitney U: U = 18 U = 12 U = 13 U = 14	.29 .00 .00 .06
11.	Weber (1968)	26	3 years	Creative Thinking Verbal Fluency	4	F = 10.58 df = 1,176	.30
12.	Thompson & Bowers (1968)	15	2 semesters	Word Meaning Social Studies	4	$F < 1$, F = 2.0 df = 1,13	.34 .46
13.	Torrance-Parent (1966)	10	2 semesters	Mathematics	7-12	rho = .32	.32

Table 4.1 Continued

	Study	No. of Teachers	Duration of Teaching	Learning Tested Subject	Grade Level	Reported Statistics	Equivalent Value in Terms of r_{xy}
14.	Allen (1970)	18	2 semesters	Arithmetic	1	p = .83;	-.23
						p = .83;	-.23
						p = .79	-.19
15.	Soar (1966)	55	2 semesters	Vocabulary; Reading	3-6	r = .068	.063
				Arithmetic (Concepts)		r = .021	.021
				Arithmetic (Problems)		r = .034	.034
				Arithmetic (Total)		r = .083	.083
						r = .081	.081
16.	Soar (1971)	35	6 months	Reading Readiness	K	r = .00	.00
		20			1	r = .30	.30
17.	Hunter (1968)	11	2 semesters	Reading, Spelling & Arithmetic	ages 8-14	r = .62	.62
18.	LaShier (1967)	10	6 weeks	Biology	8	τ = .60	.60
19.	Pinney (1969)	32	2 45-minute lessons	Social Studies & English	8-9	F = 4.2 df = 1,30	.22

divided the 26 teachers into two groups (above and below average on "indirectness") and then performed an analysis of variance _F_-test on their pupils' creative thinking test scores. Transforming the resulting _F_-ratio into an equivalent measure of _r_ took some statistical magic; hence, the form of the translation and its assumptions are characteristics of the studies that could be coded.

The point of this measurement and coding of study characteristics is to relate the properties of the studies (their subjects, investigators, technical qualities and the like) to the study findings. For example, by comparing the _r_'s for studies done at the elementary (K-6) and secondary (7-12) levels in Table 4.1, we were able to discover that the correlation between teacher indirectness and pupils' learning is higher ($\bar{r}$ = .30 based on eight cases) at the secondary level than at the elementary level ($\bar{r}$ = .16 based on ten cases), perhaps because young pupils need more direction or perhaps because lecturing style is less relevant in earlier grades (Glass _et al_., 1968).

The example of a meta-analysis of teacher indirectness was rather long; we hope that it helped make the point that the measurement of study characteristics and findings requires ingenuity and care in the definition of properties of studies and their quantification.

GENERAL CONSIDERATIONS

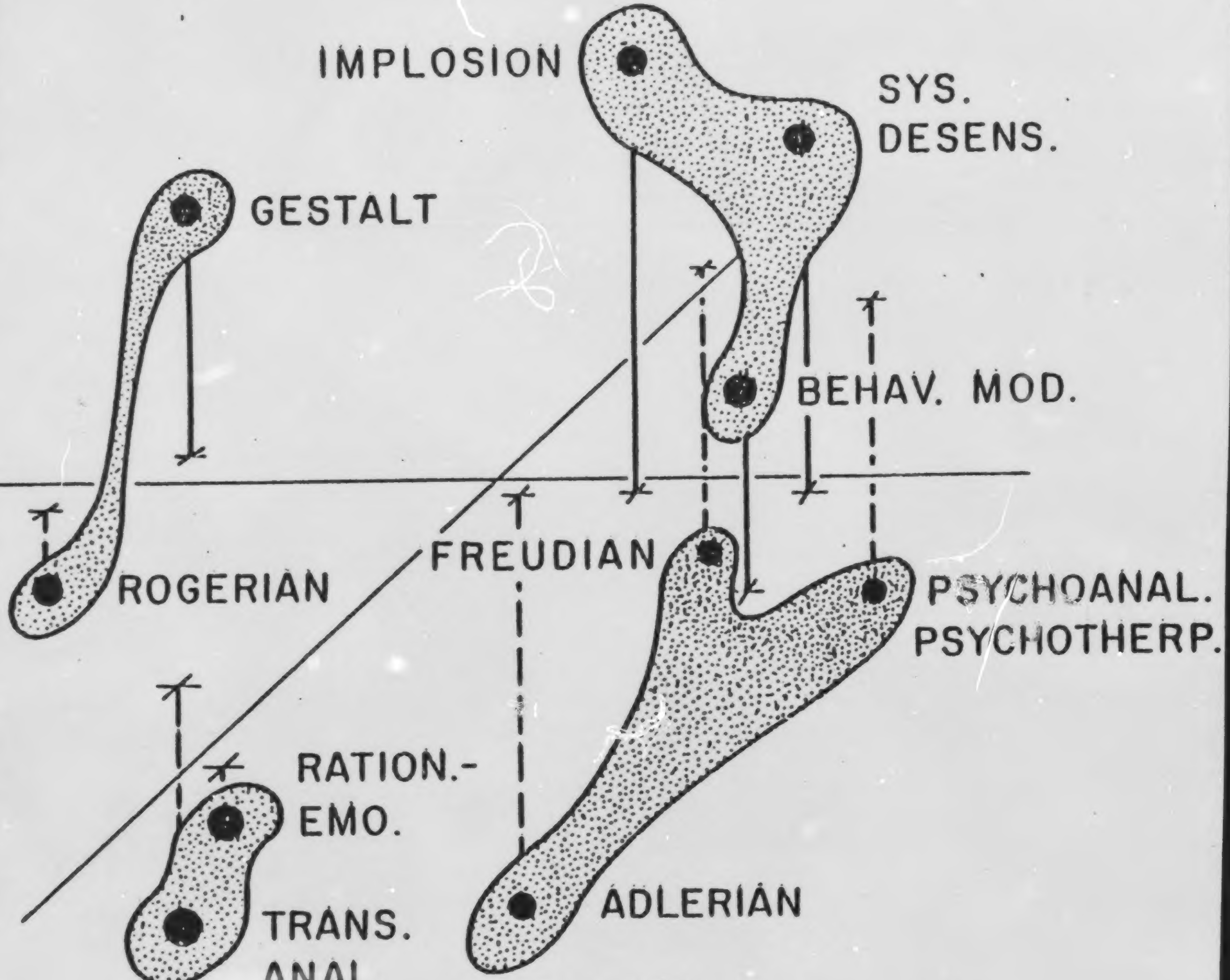
Measurement of study characteristics and findings can be evaluated with respect to both its validity and reliability, as are other instances of measurement.

Validity. The validity of measuring study properties and findings is a very broad consideration. Most things that bear on the meaning of a coded or measured characteristic are matters of validity. These considerations include such things as clarity of definitions, adequacy of reported information, the degree of inference a coder must make in determining from the written report what characterized the research, and the like. Some problems of validity can be corrected by greater care in reading and coding studies: making definitions sharper and more detailed, splitting broad concepts into more refined ones. Other problems of validity cannot easily be corrected: one must infer that in a particular study the assignment of subjects to experimental conditions was non-random because random assignment was not specified and there are significant differences on most pretest variables. There probably aren't any useful general technical guidelines for making study measurement more valid. Examples may have to substitute for principles.

Consider a somewhat extreme example of measurement of study characteristics that was pursued with more than normal care for the sake of the validity of the measurement. Smith and Glass (1977) performed a meta-analysis of nearly four hundred controlled experiments on psychotherapy outcomes. One characteristic of studies that was of principal interest was the type of psychotherapy being evaluated (e.g., Rogerian, Adlerian, behavioral, etc.). Even the simple labeling of the psychotherapy in a single study grew unexpectedly difficult at times. Could a psychotherapy described as "non-directive reflection of feeling plus empathic understanding" be properly coded as Rogerian in the absence of the investigator's having labeled it Rogerian or otherwise referred to Carl Rogers? Yes, it probably was safe to do so. But what of tougher cases? Suppose an investigator reported a study in which he compared "psychotherapy" against a wait-list control group; rather than naming the specific type of psychotherapy he merely referred to the therapists attempts "to interpret clients' defense mechanisms and help them gain insight into the causes of their difficulties." Is it safe to assume that the therapy was psychoanalytic psychotherapy and code it as such in the meta-analysis? Or would it be more prudent to classify the therapy as "eclectic insight therapy"? There's no general answer since questions at this level would be resolved by particular considerations of purposes we haven't specified. The examples merely illustrate the complexities of defining and recognizing qualities (requisites of measurement) of studies from written reports.

In our work on psychotherapy outcomes, complexities of measurement (or classification) were encountered again at a more general level. More than twenty specific types of psychotherapy appeared in the nearly 400 experiments. These twenty were fairly easily grouped into ten more general types of psychotherapy: Rogerian, Gestalt, Rational-emotive, Transactional Analysis, Adlerian, Freudian, Psychoanalytic Psychotherapy, Behavioral Modification, Systematic Desensitization, and Implosion. It was deemed worthwhile to attempt to group these ten psychotherapies into a small number of more general class so as to address additional questions in the meta-analysis. But questions remained about how this grouping might best be done. On the basis of what evidence or what process of judgment would therapies A, B and C be deemed to belong to Therapy Class I and therapies D and E to Therapy Class II? In a general sense, the question was one of measurement validity, even if measurement in this instance was only classification and coding. Perhaps the least valid grouping of therapies into homogeneous classes would have been based on our own unexplained judgment of which therapies were similar to which others. Instead, we enlisted the help of about twenty-five clinicians and counselors. For about ten hours we studied and discussed the theory and techniques of each of the ten psychotherapies. Then the therapists gave their rankings of the similarities among the psychotherapies using the method of multi-dimensional rank-ordering (Torgerson, 1958). The therapists' similarity judgments were then subjected to analysis by multi-dimensional scaling (Shepard, 1962). A graphic representation resulted of the therapists' perceptions of the similarities among the ten psychotherapies (see Figure 4.1). In the three-dimensional space in Figure 4.1, the distance between two therapies (represented by black circles) is inversely related to the similarity between the therapies in the perceptual space of the judges (therapists). The four amoeba-like figures in Figure 4.1 connect therapies that are near each other in the space. Thus, Rogerian and Gestalt therapies form a class of psychotherapies, as do Rational-emotive and Transactional Analysis. In this manner, four classes of psychotherapies were derived, and they were derived so as to reduce the influence of arbitrariness and idosyncrasy — thus, one hopes they represent a more valid classification (measurement) of studies than might otherwise have been done.

Figure 4.1. Multidimensional scaling of ten psychotherapies by 25 clinicians and counselors



Reliability. Reliability in the generic sense of the word refers to consistency of measurement. What is the extent of agreement among different measurements of the same thing? There exist many alternative ways in which the measurements to be compared for agreement may be different. For example, in the familiar instances of the reliability of measurement of human behavior the most prominent source of different measurements are temporal variations in the behavior itself. A psychologist may wish to measure peoples' mood on a scale of "happy. . . sad." He may use the same fifty-question standard inventory with each measurement so that different scores could not arise from some instability in the more mechanical aspect of the testing; but he may discover nonetheless that he obtains relatively inconsistent scores for persons because their moods are fleeting: happy in the morning, apathetic by lunch, melancholy by evening. If the psychologist chose instead to measure mood by clinical interview, the potential scources of unreliability might multiply: instability in peoples' moods across time, differences among questions posed by interviewers from one occasion to the next, differences in the standards of judgment employed by the interviewers, and the like. Cronbach and his colleagues have brought psychometrics around to the notion that the question of measurement reliability is basically the question of the relative contribution to inconsistency of measurement of multiple sources of differences among the conditions of measurement (Cronbach, Gleser, Nanda and Rajaratnam, 1971). This point of view helps one think more clearly about problems of measurement reliability that arise in research meta-analysis.

The measurement problem in meta-analysis is the problem of measuring (quantifying, classifying, coding) the characteristics and findings of studies based on written reports. That the thing measured is a written report that cannot change from one day to the next (save for spirit "ditto" copies that eventually fade into illegibility?) eliminates a major source of inconsistent measurement that plagues measurement of individual or group actions. The principal source of measurement unreliability in meta-analyses arises from different readers (coders) not seeing or judging characteristics of a study in the same way. Judge-consistency or rater-agreement is the most important consideration for our purposes.

There is no total remedy for the inconsistency that arises among different coders of the same research study. Explicit instructions, specificity in defining characteristics and Gründlichkeit will all help reduce the problem somewhat, but there are limits to what can be specified before the fact and how much detail can be imposed on coders before they quit. The guidelines we propose are 1) good sense and reasonable care at the outset, 2) assessment of the extent of disagreement by having multiple judges read a set of common studies, and 3) correction of flagrant inconsistencies discovered at step #2. Step #2 is the important one; all but the simplest meta-analyses should be subjected to an assessment of the reliability (in the rater-agreement sense of the word) of the coding procedures.

An example may help clarify this recommendation. In assessing the comparative effects of drug vs. psychotherapy, Smith, Glass and Miller (1980) developed an extensive coding system for describing the characteristics and findings of 151 experiments collected from the literature of psychopharmacology. To test the reliability of the coding, 2 judges were enlisted to code 5 studies. One judge coded 2 studies, and one coded 3. The judges were unfamiliar with the psychopharmacological literature, but well-practiced in general coding and effect-size calculation common in meta-analysis.

The 5 studies were included in the 151 studies gathered for the meta-analysis. Each judge received a drug-only study and a study of drug-plus-psychotherapy. The studies were chosen at random from all studies under ten pages in length. This restriction of length was adopted to reduce the time necessary for the judges to devote to the task. A brief list of coding conventions was given to each judge, with a request to code only the effect size for one or two dependent variables if there were many from which to choose.

One hundred sixty-two ratings were recorded by the 2 judges over the 5 studies (not including the effect sizes themselves) and were matched with an equal number of ratings by a third judge. One hundred twenty-two (75 percent) were identical and another 13 (8 percent) were within one or two scale points for five-point rating scales or continuous variables such as patient age, duration of treatment, and the like. Seventeen percent of the ratings were placed into the wrong category or were off by more than two scale points. These incorrect codings included such inconsistencies as the rating of an outcome measure as hospital adjustment rather than work adjustment or as somatic symptoms instead of anxiety. The codings of the two judges did not differ substantially from the codings of the third.

Agreement between each judge's calculation of effect sizes and an earlier independent calculation was substantial. A sixth study was added exclusively to give another test to the replicability of effect-size calculation. This study was chosen to represent a relatively complex case for calculation. Calculated by the second judge, it is reported last in Table 4.2 below.

Table 4.2 Effect sizes for two judges compared to those of a third judge

	Study	*ES* for judge	*ES* for judge no. 3	Size of error
Judge 1	Study 1	0.50	0.54	0.04
	Study 2	0.64	0.67	0.03
Judge 2	Study 3	−1.15	−0.95	0.20
	Study 4	0.87	0.85	0.02
	Study 5	1.58	1.58	0.00
	Study 6	1.08	0.93	0.15
		$\overline{ES} = 0.59$	$\overline{ES} = 0.60$	Average: 0.07

The *ES*'s, effect sizes, referred to in Table 4.2 are mean differences divided by standard deviations, a measure of experimental outcome already encountered several times in this text. It may strike the reader as curious that in only one of six instances in Table 4.2 did the two judges make calculations of effect size that agreed through two decimal places. Be assured that the discrepancies (none terribly large and on the average quite small, viz., .07) do not seem surprising at all to us. As will be seen in Chapter V, although the definition of *ES* is very simple, its calculation in particular instances can be extremely complex, frequently calling on complicated judgments about how to aggregate sources of variation, about when to make simplifying assumptions and when not to, and often entailing arduous chains of calculations in which accuracy may be compromised by rounding off a six-digit answer to four digits at some intermediate stage.

CHARACTERISTICS OF STUDIES

The characteristics of studies that are most important in a meta-analysis (apart from the findings, of course) can be roughly classified as either substantive or methodological. Substantive features are those characteristics of studies that are specific to the problem studied, e.g., in a meta-analysis of drug treatment of hyperactivity the substantive characteristics might include 1) the type of drug administered (caffeine, amphetamines, etc.), 2) the size of the dose, 3) the age of the subjects, 4) the presence or absence of checks for ingestion, and so on. The methodological characteristics of studies are more general; they may be nearly the same for all meta-analyses of a general type, such as experimental studies, correlational studies or surveys. They include a virtual table of contents of research methods books: 1) sample size, 2) test reliability, 3) randomization v. matching v. nonequivalent groups, 4) degree of subject loss, 5) single-blind, double-blind or unblinded, and the like.

The purpose underlying coding the substantive and methodological characteristics of studies is the same: one wants to learn whether the findings of the studies differ depending on certain of their characteristics. A meta-analysis seeks a full, meaningful statistical description of the findings of a collection of studies, and this goal typically entails not only a description of the findings in general but also a description of how the findings vary from one type of study to the next.

An example might clarify the use of both substantive and methodological study characteristics in this respect.

In a meta-analysis of the relationship between school class-size and pupil achievement, we coded nearly thirty substantive and methodological features of each study including the findings, viz., the standardized average difference in achievement between the larger, L, and the smaller, S, class (Glass and Smith, 1979). The characteristics coded for each study included where the study was published, in what country the research was performed, the date of publication, which subjects were taught to the pupils, and many others which can be seen in the facsimile of the coding sheet that is reproduced as Table 4.3. Using statistical models that will be presented in Chapter V, the data from over 700 comparisons of pupil achievement in smaller and larger classes were integrated into an aggregate curve descriptive of the relationship as revealed by the empirical research literature. But the analyses did not stop there. Many persons feel that the nature of the relationship between class-size and learning may vary depending on what subject is taught (math learning may flourish in small classes, but not physical education, for example) or the age of the learners. Moreover, it is possible that a flaw in research methods (unreliable tests or improper statistical analysis, for example) obscures the class-size and achievement relationship in some studies. To check for these possibilities, we analyzed the class-size and achievement relationship separately for various subdivisions of the data. For example, all studies involving pupils in grades kindergarten through six were separated from those done

Tahle 4.3

CLASS SIZE CODING SHEET

DENTIFICATION:

1) Study ID#:_____-_____.   2) Authors:__________________________.   3) Year:_____

4) Source of data: __Journal   __Book   __Thesis   __Unpublished report

5) Classification of study: __Class size   __Ability grouping   __Tutoring
__Psychol. experiment   __Secondary analysis

6) Country of origin:__________________________.

NSTRUCTION:

1) Subject taught: __Reading   __Math   __Language   __Other:____________________

2) Duration of instruction: _____hrs.   _____weeks

3) Supplemental vs. integral: __Instruction supplemented other large group instruction.
__Instruction constituted entire teaching of the subject.

4) Adaptation of instruction to class size:
Type of instruction in smaller class: ______________________________
____
Type of instruction in larger class: ______________________________
____

	Smaller Class	Larger Class
5) No. of pupils:	____________	____________
6) No. of instructional groups:	____________	____________
7) No. of instructors:	____________	____________
8) Pupil/instructor ratio:	____________	____________
9) Accuracy of estimate of ratio:	Lo Av Hi	Lo Av Hi

10) Instructor type: __Teachers   __Adult aides of tutors   __Both

11) Sex of teacher: __M   __F

12) Years teaching experience: _____years

CLASSROOM DEMOGRAPHICS:

1) Pupil ability: __IQ < 90   __90 ≤ IQ ≤ 110   __IQ > 110

2) Percent pupils female: _____%

3) Ages: 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

4) Average age: _____years

Table 4.3 (Continued)

STUDY CONDITIONS:

1) Study setting: __Regular classroom __Experimental setting

2) Assignment of Ss to groups: __Random __Matched __"Repeated measures"
__Uncontrolled

3) Assignment of instructors to groups: __Random __Matched __"Repeated measures"
__Uncontrolled

4) Percent attrition: Small class: ____% Large class: ____%

OUTCOME VARIABLE:

1) Type of Outcome Variable:
__Standardized achievement test:______________________________
__Ad hoc achievement test:______________________________
__Pupil attitude:______________________________
__Teaching behavior:______________________________
__Pupil-teacher interaction:______________________________
__Teacher attitude or satisfaction:______________________________

2) Quantification of Outcome:
__Gain scores (simple)
__Residualized gain scores
__Uncorrected dependent variable

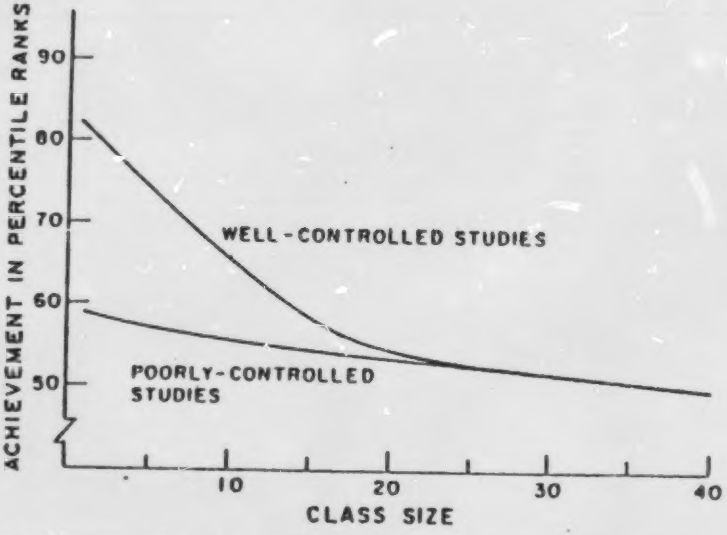
3) Congruence of instruction and outcome measure: Low Average High

4) Follow-up time: _____weeks from the end of instruction to the measurement of outcomes

5) Standardized mean difference (Small-Large):______________________________

with pupils in secondary school (a substantive characteristic). The statistical curve describing how achievement is related to class-size was then derived for each of these two parts of the data. It so happened that the two curves were nearly the same (within statistical error) so that there was no need to modify the conclusion of a class-size and achievement relationship for different age-groups of pupils.

However, one methodological characteristic of the studies was strongly related to our conclusions. Over 100 comparisons of achievement in smaller and larger classes came from studies in which the threat of pre-existing differences between classes was controlled by random assignment to the two classes; the remaining comparisons came from studies in which poor controls were achieved (e.g., naturally occurring smaller and larger classes were compared). The studies were thus distinguished with respect to a characteristic of research method. When the statistical curves were derived for these two parts of the data, quite a different picture emerged from what was seen when elementary-grade and secondary-grade studies were compared. The graphs of the two curves appear in Figure 4.2. Not only what we said about the class-size and achievement relationship but what we concluded about the trustworthiness of research on the question were affected by our discovery that the study findings varied as a function of methodological characteristics of the studies themselves.



Consistent regression lines for the regression of achievement (expressed in percentile ranks) onto class size for studies that were well-controlled and poorly-controlled in the assignment of pupils to classes.

Figure 4.2

An Example of Study Coding

In our meta analysis of psychotherapy outcome experiments (Smith, Glass and Miller, 1980), we developed a long list of substantive and methodological characteristics for describing the research literature. The numeric coding of each study extended across nearly three computer cards — 211 digits of coding in all. A facsimile of the coding sheet appears as Appendix A. It contains the following variables on which each study was classified: date of publication; form of publication; professional affiliation of the experimenter; the degree of blinding used in the study; whether more than one treatment was simultaneously compared against the control group client diagnosis; previous hospitalization; intelligence; age; sex; similarity of client to the therapist; the means by which the clients were obtained for the study; means of assigning clients and therapists to comparison groups; mortality (loss of subjects) from samples; internal validity of the study; the type, duration, modality, and location of the treatment; sample size; therapist experience; type and reactivity of outcome measure and the time after therapy when it was measured; whether factorial effects were tested; and the statistical procedures for determining the size of effect produced by the therapy. Each variable is further described below.

Each study was read and a coding form was completed for each outcome and each comparison in the study. This task presented a range of difficulty depending on the clarity of the research report and the conformity of the experimenter to standard research practices. A list of coding conventions was developed during the pilot phase of the project and was used to guide the classification of studies whose characteristics were ambiguous. These conventions are explained in the following paragraphs.

Date of Publication. This was recorded as stated on the manuscript. Some studies were published more than once, and in this case the earlist date was recorded.

Form of Publication. The study was classified according to the form in which it appeared: journal article, book, dissertation, or unpublished manuscript. If more than one form was used, such as a dissertation later published in a journal, the study was designated in its most accessible form.

Professional Affiliation of Experimenter. The study was classified, according to the affiliation of the experimenter, as either psychology, education, psychiatry, social work, or "other." This classification was determined by the institutional and departmental identification on the manuscript, or by membership in the American Psychological Association.

Blinding of Experimenter. This variable represents the degree of blinding that prevails in the assessment of outcomes or in the administration of these in the study. If the experimenter or the outcome evaluator was kept uninformed about whether each subject was in the control group or the treated group, the study was classified as "single blind." If no information was provided that showed that the experimenter or evaluator was kept uninformed about group composition, the study was categorized as either "experimenter did the therapy" or "experimenter knew the composition of the groups but didn't personally treat the client."

Client Diagnosis. In the meta-analysis, the diagnostic label that the experimenter used was recorded and classified into a twelve-category diagnostic system. The categories were (1) neurotic or true (complex) phobic, (2) simple (monosymptomatic) phobic, (3) psychotic, (4) normal, (5) character disordered, (6) delinquent or felon, (7) habituee (e.g., alcohol, tobacco, drug addiction), (8) emotional-somatic disordered, (9) handicapped (physically or mentally), (10) depressive, (11) mixed diagnoses, and (12) unknown.

Hospitalization. The number of years of previous hospitalization, as stated or implied by the author, was another indication of the severity of client distress and was recorded.

Intelligence. Intelligence of the client is frequently cited as mediating the effects of psychotherapy. The intelligence of the group was rated as "below average" for IQ scores less than 95, "average" for IQ scores between 95 and 105, and "above average" for IQ scores above 105. The source of information about client intelligence was also recorded. In 4 percent of the studies, IQ was reported by the experimenter. In 61 percent of the studies, IQ could be inferred (at least with the accuracy necessary to make the three gross distinctions) from the client's placement in some institution, such as a college or a treatment facility for the mentally retarded. In 35 percent of the cases, client intelligence could not be assessed from the report and therefore was estimated as average.

Client-Therapist Similarity. The socioeconomic and ethnic similarity between client and therapist is also thought to influence the outcome of therapy. The cultures of the therapist and the client are similar in the sense that they share common languages, value systems, and educational backgrounds. The more healthy the client, the more he resembles the therapist. The studies were rated for similarity between the client and the typical white, middle-class, well-educated therapist. The highest value (4) was used for studies of white, middle-class, well-educated, and mildly or moderately distressed clients. The lowest value (1) was used when the typical therapist treated lower-class minority or severely disturbed clients.

Solicitation of Clients. The use of volunteers in therapy studies has been sufficient cause for some previous reviewers to disallow these studies as tests of therapeutic effect. Yet in the case of most analogue studies, the volunteers reported symptoms, requested and were given psychological treatments to remedy them. It is possible that they differ only in degree from "real" clients who independently seek treatment. The studies were classified according to whether (1) the subjects were solicited for therapy by the experimenter (usually by offering treatment to psychology students who obtained extreme scores on anxiety measures); (2) the subjects came to the treatment program in response to an advertisement; (3) the subjects recognized the existence of a problem and sought treatment; (4) the subjects were referred for treatment; or (5) the subjects were committed to the treatment, with no choice.

Assignment to Groups. A characteristic often afforded most importance in judging the validity of a comparative study is how the experimenter allocated subjects to treated and control groups. Random assignment insures, within probability limits, that the two groups are initially comparable and that differences between them on the post-test are attributable either to chance (with probability equal to the significance level) or to the treatment and to no other source of influence. Matching pairs of subjects is the next best method, although using it presumes that all sources of influence on therapy are known and can be used as matching variables.

Moreover, it renders significance levels meaningless when calculated in the usual ways. Ex post facto matching, covariance adjustments, and equating on pretest scores are less satisfactory allocation methods, but still better than no matching at all. Studies were classified according to the assignment of both clients and therapists to groups.

Experimental Mortality. Dropouts from treatment and control groups represent a critical problem in psychotherapy research. Eysen ck and Rachman declared that a dropout must be considered a treatment failure. Yet early termination can be explained by a variety of reasons other than treatment failure. These include economic problems, family or work problems unconnected with the psychological difficulties, amelioration of symptoms, scheduling changes, physical illness unrelated to treatment, and even death. Unless these alternative explanations are accounted for, the premature terminators cannot be classified as either successes or failures. Yet the decision to include or exclude terminators from final statistics may have a substantial effect on the findings of a study. Because the decision is made on professional judgment rather than independent empirical justification, the decision invites bias.

Premature termination is best regarded as a problem of the internal validity of the study and not confounded with outcome measurement. In this study, the percent mortality was coded separately for treated and untreated groups. These figures were occasionally difficult to ascertain and involved comparing degrees of freedom in post-test analyses with the numbers of subjects orginally allocated to groups. A study might also have different rates of mortality at the times of the post-test and the follow-up. These different mortality percents were noted separately.

Internal Validity. The internal validity of a study was judged on the basis of the assignment of subjects to treatment and the extent of experimental mortality in the study. To be judged high on the internal validity scale, a study must have used random assignment of subjects to groups and have a rate of mortality less than 15 percent and equivalent between the two groups. If mortality was higher or nonequivalent, internal validity was still rated high if the experimenter included the scores of the terminators in the post-test statistics or established the initial equivalence of terminators and nonterminators. Medium internal validity ratings were given to (1) studies with randomization but high or differential mortality; (2) studies with "failed" randomization procedures (e.g., where the experimenter began by randomizing, but then resorted to other allocation methods, such as taking the last ten clients and putting them into the control group) with low mortality; and (3) extremely well-designed matching studies. Low validity studies were those whose matching procedures were quite weak or nonexistent (e.g., where intact convenience samples were used) or where mortality was severely disproportionate. Occasionally, statistical or measurement irregularities decreased the value assigned to internal validity, such as when an otherwise well-designed study employed different testing times for treated and untreated groups. This measure of internal validity was not contaminated by sample size, reactivity of measures, or the degree of blinding employed in the study. All four constructs were assessed separately.

Allegiance of the Experimenter. Faith in the therapy on the part of the therapist has been mentioned as a putative cause of positive therapeutic effects. From the tone and substance of the research report, it was usually possible to determine whether the experimenter was partial to the treatment evaluated. For example, when the report contained enthusiastic endorsements of the therapy, this variable was coded as positive. Where a second therapy was clearly a foil for the favored therapy, this variable was coded as negative. Placebo treatments were always coded as negative. Where the experimenter was the therapist, this variable was coded positive.

Therapy Modality. Each study was coded for the modality in which the therapy was delivered -- individual, group, family, mixed modalities, automated, or "other."

Treatment Location. Each study was coded according to the location in which the therapy was delivered -- school, hospital, mental health center, other clinic, private practice, college facility, prison, residential facility, or "other."

Therapy Duration. The duration of therapy, both in number of hours and weeks, was recorded. The rate (hours per week) of therapy was computed from these two variables.

Therapist Experience. The number of therapists used in the study and their experience in years was recorded. Because reports were frequently lacking this information, the following conventions were developed for translating relevant bits of information into years of therapist experience when no more specific information was given:

Undergraduates or other untrained assistants = 0 years
MA candidates = 1 year
MA-level counselor or therapist = 2 years
Ph.D. candidate or psychiatric resident = 3 years
Ph.D.-level therapist = 5 years
Well-known, Ph.D.-level therapist = 7+ years

Outcome Measurements. Previous reviewers have struggled with the philosophical and technical problems connected with the selection and measurement of outcomes. A reviewer might count a study as supportive or not supportive of the effectiveness of psychotherapy based on the statistical significance of the outcome measure. Yet most studies employed more than one outcome measure, using different instruments or the same instrument given at different times after therapy. When different measures produced different results, several strategies were employed to cope with this problem. A study could be counted twice, for example, with one vote for and one against the therapy. Or, if a study showed positive effects at therapy termination, but no effects at the follow-up, that study could be listed as a negative indicator of therapeutic effectiveness. This strategy exemplifies a confusion between the use of empirical research for theory building and research done for evaluative purposes, i.e., to determine the effects and practical value of a treatment. The direction of desirable therapeutic effect was obvious in nine out of ten cases by examining the research hypotheses stated by the experimenter or the narrative description of results. In the remainder of cases, the Mental Measurements Yearbooks were consulted, or other studies that had employed the same measure. Each outcome measurement listed by the experimenter was used in the meta-analysis. Each measure was weighed equally; however, redundant measures were eliminated. If, for example, a second measure matched the first in outcome type, degree of reactivity, follow-up time, and approximate size of effect, the second measure was deemed redundant and ordinarily not included in the meta-analysis. When subtest scores of multifactorial test batteries (e.g., MMPI) were reported, and the subtests yielded results that were only randomly different from one another, an average of the subtests was used. Total test battery results were used in favor of separate subtest scores.

The specific outcome was recorded and grouped into one of twelve outcome types: (1) fear or anxiety measures; (2) measures of self-esteem; (3) tests and ratings of global adjustment; (4) life indicators of adjustment; (5) personality traits; (6) measures of emotional-somatic disorders; (7) measures of addiction; (8) sociopathic behaviors; (9) social behaviors; (10) measures of work or school achievement; (11) measures of vocational or personal development; and (12) physiological measures of stress. The table below contains the outcome measures that were grouped within two outcome types: life indicators of adjustment and social behaviors:

Outcome labels grouped into two outcome types

Outcome type	
Life indicators of adjustment	Social behaviors
Number of times hospitalized	Interpersonal maturity
Length of hospitalizations	Interpersonal interaction
Time out of hospital	Social relations
Employment	Assertiveness
Discharge from hospital	IPAT sociability scale
Completion of tour of duty	Acceptance of others
Recidivism	FIRO-B
	Dating behavior measures
	Problem behavior in school social setting
	Social effectiveness
	Social distress
	Sociometric status
	Social distance scale
	Social adjustment

Reactivity of Outcome Measure. Highly *reactive* instruments are those that reveal or closely parallel the obvious goals or valued outcomes of the therapist or experimenter; which are under control of the therapist, who has an acknowledged interest in achieving predetermined goals; or which are subject to the client's need and ability to alter his scores to show more or less change than what actually took place. Relatively nonreactive measures are not so easily influenced in any direction by any of the parties involved. Using this definition of reactivity, it was possible to define a five-point scale with the low end anchored at unreactive measures, such as physiological measures of stress (e.g., Palmar Sweat Index) and anchored at the high end with therapist judgments of client improvement. Points on the scale are further illustrated in the following table:

Conventions for assigning values of reactivity to tests and ratings

Reactivity value	Tests and ratings of therapy outcome
1 (lowest)	Physiological measures (PSI, Pulse, GSR), grade point average
2	Blinded ratings and decisions—blind projective test ratings, blind ratings of symptoms, blind discharge from hospital
3	Standardized measures of traits having minimal connection with treatment or therapist (MMPI, Rotter I-E)
4	Experimenter-constructed inventories (nonblind), rating of symptoms (nonblind), any client self-report to experimenter, blind administration of Behavioral Approach Tests
5 (highest)	Therapist rating of improvement or symptoms, projective tests (nonblind), behavior in the presence of therapist or nonblind evaluator (e.g., Behavioral Approach Test), instruments that have a direct and obvious relationship with treatment (e.g., where desensitization hierarchy items were taken directly from measuring instrument)

Treatment. To determine whether the therapeutic effect produced in a study was related to the type of treatment used, a system for categorizing treatments was developed.

1) Psychodynamic therapies were those employing concepts such as unconscious motivation, transference relationship, defense mechanisms, structural elements or personality (id, ego, superego) ego development and analysis.

2) Dynamic-eclectic therapies are based on dynamic personality theories, but employ a wider range of therapuetic techniques and interactive concepts than the more orthodox Freudian theory.

3) Adlerian therapy (Adler is referenced by Dreikurs and others) is based on the never-ending strivings of the personality to escape from a sense of inferiority. Striving for superiority alienates people from love, logic, community life, and social responsibility.

4) Hypnotherapy (Wolberg) is one type of therapy that uses hypnosis as a tool for increasing relaxation and suggestibility and weakening ego defenses. As described by Lewis Wolberg, hypnotherapy is closed related to psychodynamic theory, suggesting that such neurotic states as anxiety, hysteria, and compulsions are susceptible to this treatment.

5) Client-centered or nondirective psychotherapy is associated with Rogers, Truax, Carkhuff, Gendlin, and Axline (nondirective play therapy with children) among others. The key concepts of this therapy include the necessary conditions of therapist congruence, empathy, and unconditional positive regard

for the client.

6) Gestalt therapy was developed by Perls (Perls, Hefferline, and Goodman) and, like Rogerian therapy, is humanistic and phenomenological in philosophy. The key concept in this therapy is awareness. The healthy person can readily bring into awareness all parts of his personality and apprehend them as an integrated whole. Therapy is a process of heightening awareness through immediate here-and-now emotional and physical experiences and exercises and integrating alienated elements in the person (e.g., healing the "splits" between body and mind, conscious and unconscious).

(7) Rational-emotive psychotherapy was developed by Ellis and rests on a cognitive theory of human personality and therapeutic intervention. The ABC theory holds that human reactions (C) follow from cognitions, ideas, and beliefs (B) about an event, rather than from the event itself (A). The beliefs may be either rational (logical, empirical) or irrational. These irrational beliefs are common for people in distress and pervasive in our society. They include the notion that one must be universally loved, or that failure at a task is utterly catastrophic. The therapist demonstrates the ABC theory in relation to the client's problem, convinces the client of the truth of the theory, confronts the irrational reactions, and teaches the client to confront them himself. The objective of therapy is to replace the irrational, self-defeating cognitions with logical and empirically valid cognitions.

8) Other cognitive therapies comprise a family of therapeutic theories related to Ellis's rational-emotive psychotherapy in that the place of cognitive process — faulty beliefs, irrational ideas, logically inconsistent concepts — is central. Theorists in this family include George Kelly, Victor Raimy, and Donald Tosi. They are similar in that the therapies are often active, didactic, directive, sometimes bordering on being hortatory. The therapists confront logical inconsistencies, interpret faulty generalizations and self-defeating behaviors, assign tasks to work on, and generally use suggestion and persuasion to get the client to give up his self-defeating belief system.

9) Transactional analysis is primarily associated with Eric Berne who developed a personality theory based on three ego states — the parent, adult, and child — and the interrelationship of these ego states within a person and between persons. All beliefs, cognitions, and behaviors are under the control of these ego states. Therapy consists of on-going (usually group) diagnosis and interpretation of the structural elements of communication and interaction, with the goal of improved reality testing and complementary transactions.

10) Reality therapy is identified with William Glasser and is based on the idea that persons who deny reality are unsuccessful and distressed. Mental illness does not exist -- only misbehavior that is based on the denial of reality. Reality is achieved by the fulfillment of the basic needs -- to love and

and be loved and to feel self-worth (success identity). The therapist establishes a personal relationship with the client; attends to present behavior rather than historical events or feelings; interprets behavior in light of the theory; encourages the formation of value judgments about correct behavior and a plan for changing behavior, rejecting excuses for a failure to change, and the development of self-discipline.

11) Systematic desensitization is a therapy based on scientific behaviorism, primarily associated with Wolpe. In this therapy, anxieties are eliminated by the contiguous pairing of an aversive stimulus with a strong anxiety-competing or anxiety-antagonistic response. The usual procedure is to teach the client deep muscle relaxation (a response antagonistic to anxiety) and then introduce anxiety-provoking stimuli, arranged in hierarchies, in connection with the relaxation until the client can confront and overcome the anxiety directly. The behavioral principles involved are reciprocal inhibition, counter-conditioning, or extinction.

12) Implosive therapy, developed by Stampfl, operates on many problems similar to those addressed by systematic desensitization, and is based on classical conditioning models. The therapist directs the client's imagery so that he is forced to imagine the worst possible manifestation of his fear, and the connection between conditioned stimulus and conditioned response is extinguished.

13) Operant-respondent behavior therapies are a family of treatment programs in which the scientific laws of learning are invoked. The client is viewed as a passive recipient of reinforcement or conditioning. Proponents include Skinner, Staats, Bijou, and Baer.

14) Cognitive behavior therapies are a family of therapies in which laws of learning are applied to cognitive processes. Unlike the strictly operant or respondent theories, in cognitive-behavioral therapies, the client is more of an active agent in his own therapy, occasionally even administering the treatment himself (e.g., self-control desensitization). Modeling treatments are included in this family of therapies because the client must identify with the model and adopt the behavior for which the model (but not the client) is reinforced. Among the proponents of cognitive behaviorism are Donald Meichenbaum, Albert Bandura, and Mahoney.

15) Eclectic-behavioral therapy is a collection of treatments that employ behavioral principles in training programs designed to affect a variety of emotional and behavioral variables. Assertiveness training is the principal therapy, and Lazarus and Phillips are among the proponents.

16) Vocational-personal development counseling involves providing skills and knowledge to clients to facilitate adaptive development. Frequently, a trait and factor approach is used with aptitude and personality testing, diagnosis, prescription, and

interaction with the client to facilitate the development of personal, social, educational, and vocational skills. Among the proponents are Theo Volsky and Williamson.

17) "Undifferentiated counseling" refers to therapy or counseling that lacks descriptive information and references that would identify it with proponents of theory. It is usually practiced in schools (i.e., the clients were given ordinary counseling), but sometimes is used as a foil against which a more highly valued therapy can be compared. That it cannot be attributed to any single theorist or group of writers is indicative of its lack of theoretical explication.

18) Placebo treatments were often included in an experimental study of therapeutic effectiveness. Placebos were used to test the effects of client expectancies, therapist attention, and other nonspecific and informal therapeutic effects. The placebo treatments tested in the meta-analysis were the following: relaxation training, attention control, relaxation and suggestion, relaxation and visualization of scenes in an anxiety hierarchy, group discussion, reading and discussing a play, informational meetings, pseudo-desensitization placebo, written information about the phobic object, bibliotherapy, high expectancy placebo, visualization of reinforcing scenes, minimal contact counseling, T-scope therapy, pseudo-treatment control, and lectures.

A scale was developed to indicate the degree of confidence in classifying therapy labels into therapy types. The greater the number of concepts, descriptions, and proponents named by the experimenter and associated with a major school of thought, the higher the value assigned to this scale. The highest value (5) was given to a study when the major proponent of a theory actually participated in the study, or when the therapy sessions were recorded and rated for their fit with the theory. The low point of the scale (1) was given to studies when the experimenter provided almost no key concepts or references. On this five-point scale, 15 percent of the studies fell into the highest category, 42 percent in the next highest, 24 percent in the middle category, and 19 percent in the lowest two categories. The mean for the confidence of classification scale was 3.5 (standard deviation = 1.0).

We have presented so much detail about the psychotherapy study characteristics and the conventions for coding because we can imagine that many of the items, particularly those dealing with experimental methods, are of general usefulness. This chapter concludes with an example of a study coded according to the conventions described above and the items on the psychotherapy study coding form in Appendix A. The study used as an example was performed by Krumboltz and Thoresen (1964) and is reproduced in Appendix B. Its description appears as Table 4.4.

Table 4.4

Classification of a study by Krumboltz and Thoresen (1964)

Publication date	1964
Publication form	Journal
Training of experimenter	Education (known by institutional affiliation)
Blinding	Experimenter (evaluators) did not do therapy, but did know group composition (no information about blinding of evaluators was given)
Diagnosis	Vocationally undecided (students who asked for counseling about future plans, grouped in "neurotic" diagnostic type)
Hospitalization	None
Intelligence	Average (estimated, in the absence of other information)
Client-therapist similarity	Moderately similar (ages differed, but socioeconomic status of community indicated similarity)
Age	16 (high school juniors)
Percentage male	50% (sample stratified by client sex)
Solicitation of clients	Clients volunteered after being given notice that counseling would be available
Assignment of client	Random (stated)
Assignment of therapist	Random
Experimental mortality	No subjects lost from any group (stated)
Internal validity	High
Simultaneous comparison	Yes (2 treatments groups and placebo group compared against control)
Type of treatment	(1) Model reinforcement—Cognitive behavioral subclass (students were shown tapes of models being reinforced for information-seeking behavior, but students were not reinforced personally) (2) Verbal reinforcement—Behavioral subclass (counselors verbally reinforced clients for production of information-seeking statements) (3) Film discussion—Placebo (clients saw and discussed a film, to control for nonspecific effects of counselor attention)
Confidence of classification	Rated 5 (highest) (because of thoroughness of description, knowledge of experimenters' theory and previous work)
Allegiance	Equal allegiance paid to each of treatments. No allegiance to placebo condition
Modality	Mixed (students were randomly assigned to individual and group treatments, but modality did not interact with outcome, so the two modes were combined for the meta-analysis)
Location	School (stated)
Duration	2 hours, 2 weeks (2 sessions, time estimated)
Experience of therapists	2 years (estimated by status in counselor-training program plus training for this experiment)
Outcome	Two outcome measures were used: frequency and variety of information-seeking behavior as estimated from responses to structured interview questions. Reactivity was rated "4" for both, because measures were self-report of clients to nonblind evaluators. These were classified as measures of vocational or personal development
Effect size	Statistics reported as treatment means and mean squares from a 4-factor analysis of variance

The effect sizes were as follows:

	Frequency (of information-seeking behavior)	Variety (of information-seeking behavior)
Model reinforcement	1.29	0.77
Verbal reinforcement	1.05	1.39
Placebo	0.21	0.27

CHAPTER FIVE

MEASURING STUDY FINDINGS

All quantitative, empirical studies aim to assess a particular phenomenon. In the case of experiments, that phenomenon is an effect of an independent variable on a dependent variable and it is measured by a difference between means, perhaps more than one such difference from a single experiment. In the case of correlational studies, the phenomenon of principal interest is the relationship between two variables, its strength and direction, usually expressed on a scale derivative of Pearson's notion of product-moments. In surveys, attention often focuses on a simple rate or incidence figure, e.g., 37 percent of people live in multiple-family dwellings. In a meta-analysis, it is the findings of studies that correspond to the dependent variable. They are to be measured in quantitative and comparable terms, then described and accounted for by reference to the "independent" and "mediating" variables that are the study characteristics discussed in Chapter Four.

In this chapter, we shall first consider the crudest level of quantification of study findings, a level that is typical of recent techniques of research study integration. At this first level, studies are classified only as "statistically significant" of "nonsignificant." This primitive translation of complex findings into crude categories proves to have some unexpected drawbacks; and in modified forms, it may yet prove to have some advantages in a few special instances. Then we shall discuss at length the properties and uses of Δ_{E-C}, the standardized mean difference for describing experimental effects. A special aspect of this problem that will be addressed is the measurement of experimental effects, Δ, for dichotmously measured outcome variables. A brief section will be devoted to the measurement of findings in correlational studies. The chapter concludes with a description of a measure of effect size recently proposed by Kraemer and Andrews.

Vote-Counting and Other Crude Measures of Study Findings

The most commonly used method of integrating research studies is what Light and Smith (1971) referred to as the voting method. There exists a virtually huge number of such reviews, and no purpose would be served by citing examples here. Light and Smith characterized the voting method in these words:

> All studies which have data on a dependent variable and a specific independent variable of interest are examined. Three possible outcomes are defined. The relationship between the independent variable and the dependent variable is either significantly positive, significantly negative, or there is no significant relationship in either direction. The number of studies falling into each of these three categories is then simply tallied. If a plurality of studies falls into any one of these three categories, with fewer falling into the other two, the modal category is declared the winner. This modal categorization is then assumed to give the best estimate of the direction of the true relationship between the independent and dependent variable. (p. 443)

Light and Smith pointed out that the voting method of study integration disregards sample size. Large samples produce more "statistically significant" findings than small samples. Suppose that nine small-sample studies yield not quite significant results, and the tenth large-sample result is significant. The vote is one "for" and nine "against," a conclusion quite at odds with one's best instincts. So much the worse for the voting method. Precisely what weight to assign to each study in an aggregation is an extremely complex question, one that is not answered adequately by suggestions to pool the raw data (which are rarely available) or to give each study equal weight, regardless of sample size. If one is aggregating arithmetic means, a weighting of results from each study according to $\sqrt{n}$ might make sense, reasoning from an admittedly weak analogy between integrating study findings and combining independent random samples from a population. The problems of proper integration of statistical findings are not simply problems of sample size; if pursued for long, they lead back to the ambiguities of the concept of a "study."

Some of the complications of sample size can be avoided post hoc if the sample size, n, of studies is not systematically related to the magnitude of the findings of the studies, for example, mean differences or correlation coefficients. Glass and Smith (1976) found for over 800 measures of the experimental effect of psychtherapy *versus* a control condition that the effect size had a linear correlation of only -- .10 with n and essentially no curvilinear correlation. Smaller size studies tended to show slightly larger effects, but the relationship was so weak that it is doubtful that any weighting of findings would make any difference in the aggregation.

A serious deficiency of the voting method of research integration is that it discards good descriptive information. To know that televised instruction beats traditional classroom instruction in 25 of 30 studies -- if, in fact, it does -- is not to know whether TV wins by a nose or in a walkaway. One ought to integrate measures of the strength of experimental effects or relationships among variables (according to whether the problem is basically experimental or correlational). Researchers commonly believe that significance levels are more informative than they are. Tallies of statistical significance or insignificance tell little about the strength or importance of a relationship.

An example will demonstrate that the aggregation of even simple statistical information can create unexpected difficulties. There exists a paradox attributed to E. H. Simpson by Colin Blyth (1972) which has a counterpart in aggregating research results. Imagine that researcher A is conducting a study of the effect of amphetamines on hyperactivity in sixth-grade children. (It is alleged that amphetamines act as depressants on prepubescent children.) In A's study, 110 hyperactive children receive the amphetamine, and 70 receive a placebo. After six weeks' treatment, each child is rated as either 'improved' or 'worse'. The following findings are obtained:

	Study A		
	Amphetamine	Placebo	
Improved	50	30	80
Worse	60	40	100
	110	70	180

The improvement rate for the amphetamines exceeds that for the placebo: .45 vs. .43.

Suppose researcher B is studying the same problem at a different site and obtains the following results:

Study B

	Amphetamine	Placebo	
Improved	60	90	150
Worse	30	50	80
	90	140	230

Again, the improvement rate for amphetamines is superior to that for the placebo: .67 vs. .64.

By the voting method of aggregation, the score would be 2-0 in favor of amphetamines. However, an aggregation of the raw data produces the opposite conclusion:

Studies A & B Combined

	Amphetamine	Placebo	
Improved	110	120	230
Worse	90	90	180
	200	210	410

The improvement rate for placebo now exceeds that for amphetamines: .55 for amphetamines vs. .57 for placebo.

Which method of aggregation is correct? Obviously they cannot both be correct, since they lead to contradictory conclusions. In pondering this paradox and its implications for research integration, it is helpful to note that (1) the paradox has nothing whatever to do with statistical significance, (2) the sizes of the differences in rates could be made as large or small as one wished by juggling the figures, (3) the basic problem is related to the problems of unbalanced experimental designs (Simpson's paradox could not occur if amphetamine and placebo groups were of equal size within each study), and (4) the practical consequences of the paradox are not negligible -- it occurred, for example, in a study of sex bias in graduate school admissions (see Bickel, Hammel, & O'Connell, 1975; Gardner, 1976).

Hedges and Olkin (1980) discovered some intriguing and unexpected deficiencies in the vote-counting method of integrating studies. They assumed that J studies each with sample size n are performed. In each study, the same effect size $\Delta = (\mu_E - \mu_C)/\sigma_C$ is estimated. The findings of each study are evaluated by a two-tailed t-test of mean differences at the .05 level of significance. Each result is classified into one of three categories: negative significant, positive significant, or statistically insignificant. The decision rule is that the over-all result is regarded as supporting the hypothesis (that μ_E is greater than μ_C) if a plurality (i.e., greater than one-third) of the studies fall into the "positive significant" category.

Hedges and Olkin assumed normally distributed variables and then calculated the probabilities for various sample sizes and numbers of studies, J, that more than a third of the studies would fall in the "positive significant" category. In Table 5.1 appear the one's complement of these probabilities; thus, the tabulated probability is the probability of failing to detect an effect size, Δ, of a given size by the one-third plurality rule. Consider, for example, the case

TABLE 5.1

Probability that a Standard Vote Count Fails to Detect an Effect for Various Sample, Effect and Cluster Sizes. Each of the $\mathcal{J}$ replicated studies has a common sample size n. A two-tailed t-test is used to test mean differences at the .05 level of significance. An effect is detected if the proportion of positive significant results exceeds one-third.

Number, $\mathcal{J}$, of studies to be integrated	Sample size, n, per study	Effect size $\Delta \equiv (\mu_E - u_c)/\sigma_c$							
		.1	.2	.3	.4	.5	.6	.7	.8
10	10	1.00	.999	.998	.994	.985	.968	.935	.880
10	20	1.00	.998	.990	.966	.906	.987	.606	.395
10	30	.999	.995	.975	.906	.947	.502	.252	.089
10	40	.999	.991	.950	.813	.547	.254	.073	.012
10	50	.999	.986	.914	.694	.358	.105	.016	.001
15	10	1.00	1.00	1.00	.999	.997	.991	.975	.939
15	20	1.00	1.00	.999	.991	.958	.862	.672	.419
15	30	1.00	.999	.994	.958	.824	.549	.244	.064
15	40	1.00	.999	.983	.885	.604	.246	.049	.004
15	50	.999	.997	.962	.770	.373	.080	.006	.000
20	10	1.00	1.00	1.00	.999	.997	.989	.966	.914
20	20	1.00	1.00	.998	.988	.941	.800	.545	.265
20	30	1.00	1.00	.993	.941	.747	.400	.118	.016
20	40	1.00	.999	.978	.834	.463	.119	.011	.000
20	50	1.00	.997	.948	.672	.222	.023	.001	.000
25	10	1.00	1.00	1.00	1.00	.999	.997	.986	.954
25	20	1.00	1.00	1.00	.996	.971	.863	.610	.291
25	30	1.00	1.00	.998	.972	.815	.448	.120	.013
25	40	1.00	1.00	.992	.892	.519	.121	.008	.000
50	10	1.00	1.00	1.00	1.00	1.00	1.00	.998	.986
50	20	1.00	1.00	1.00	1.00	.994	.915	.589	.174
50	30	1.00	1.00	1.00	.994	.862	.363	.035	.000
50	40	1.00	1.00	.999	.942	.461	.036	.000	.000
50	50	1.00	1.00	.995	.773	.124	.001	.000	.000

of 15 studies in each of which Δ is estimated from $n = 50$ cases and the true effect size being estimated equals .40, a fairly large effect. Hedges and Olkin's table shows that the probability of _not_ deciding that there is a positive effect using the vote-counting strategy is .770, i.e., the probability of error is greater than three-quarters! What is even more remarkable is that for $\Delta < .40$, the probability of making the error indicated _increases_ as the number of studies integrated increases. Clearly, there is much that is unacceptable in research integration by means of vote-counting.

Integrating Significance Tests

Some researchers have set forward as the principal problem of research integration the combining of significance levels into a joint test of a null hypothesis. Gage (1976) contributed a considered and illuminating paper on integrating studies on teaching. Following an astute critique of the voting method, he posed the aggregation problem as a problem in determining whether several individual studies, many of which showed no significant correlation, constituted in the aggregate sufficient evidence to reject the null hypothesis at a high level of significance. He employed the chi square method of K. Pearson (1933) and E. S. Pearson (1938) via Jones and Fiske (1953). If k independent studies yield significance levels, $p_1, p_2, \ldots, p_k$, then under the common null hypothesis tested in each study:

$$-2\sum_{i=1}^{k}\log_e p_i \sim \chi^2_{2k}$$

This approach seems defensible and more powerful than a binomial test -- testing whether the probability of "positive" findings is different from .5 -- where statistical hypothesis testing is a genuine concern. For most problems of meta-analysis, however, the number of studies will be so large and will encompass so many hundreds of subjects that null hypotheses will be rejected routinely. Perhaps it is more realistic to think of the typical meta-analysis problem as residing in that vicinity the statistician calls "the limit," where all null hypotheses are false and inferential questions disappear. The statistical integration of studies probably ought to fulfill descriptive purposes more than inferential ones, though obviously it may fulfill both.

If the Pearson χ^2 test of combined results begins to play an increasingly important role in research integration, methodologists will need to scrutinize its assumptions and properties. It is probably quite sensitive to nonindependence of studies (cf. Jones & Fiske, 1953, pp. 317-381). Furthermore, the extreme tails of distributions are exotic places about which more would have to be learned. For example, violation of normality assumptions has little effect on 95th and 99th percentiles of t and F distributions, but conceivably it can change a p of .001, under normality, to a p of .0001, which is a disturbance in natural logarithms from -6.91 to -9.21.

Rosenthal (1978) recently evaluated nine different methods that have been used at one time or another to aggregate statistical significance measures from many studies. These methods include addition of logs of p-levels mentioned above as well as adding probabilities

(Edginton, 1972a), adding *t*'s (Winer, 1971), Stouffer's method of adding *Z*'s (Mosteller and Bush, 1954), adding weighted *Z*'s (Mosteller and Bush, 1954), testing the average *p*-level (Edgington, 1972b), testing the average *Z* (Mosteller and Bush, 1954), counting (vote-method) and blocking (see Rosenthal, 1978, p. 190). Rosenthal's summary of the advantages and limitations of the various methods appears as Table 5.2.

Table 5.2*

Advantages and Limitations of Nine Methods of Combining Probabilities

Method	Advantages	Limitations	Applicable when
Adding logs	Well established	Cumulates poorly; can support opposite conclusions	*N* of studies is small (≤ 5)
Adding *p*s	Good power	Inapplicable when *N* of studies (or *p*s) is large, unless complex corrections are introduced	*N* of studies is small ($\Sigma p \leq 1.0$)
Adding *t*s	Unaffected by *N* of studies, given minimum *df* per study	Inapplicable when *t*s are based on very few *df*	Studies are not based on too few *df*
Adding *Z*s	Routinely applicable, simple	Assumes unit variance when under some conditions Type I or Type II errors may be increased	Anytime
Adding weighted *Z*s	Routinely applicable, permits weighting	Assumes unit variance when under some conditions Type I or Type II errors may be increased	Whenever weighting is desired
Testing mean *p*	Simple	*N* of studies should not be less than four.	*N* of studies ≥ 4
Testing mean *Z*	No assumption of unit variance	Low power when *N* of studies is small.	*N* of studies ≥ 5
Counting	Simple and robust	Large *N* of studies is needed; may be low in power.	*N* of studies is large
Blocking	Displays all means for inspection, thus facilitating search for moderator variables	Laborious when *N* is large; insufficient data may be available.	*N* of studies is not too large

---

*After Rosenthal (1978), reprinted by permission of the author and publisher.

Scaling Experimental Findings

For several reasons and in several ways it may occur that the findings of a comparative study exist only in the form of a report whether one mean (median or whatever) is higher or lower than another. This most basic report of a finding can arise from 1) very rudimentary reporting in a brief article, 2) the desire to avoid making dubious assumptions, or 3) incomplete data which obviate the calculation of a metric measure of effect or correlation. Thus, a data analyst attempting to integrate the findings of many studies may have in hand data of the following type: in 75 comparisons of treatments A and B, A exceeded B 45 times on the outcome measure, and B exceeded A the other 30 times. The key to converting these rudimentary results into metric measures of effects or correlation lies in traditional methods of psychometric scaling. In particular, if one can assume normality, then Thurstone's "law of comparative judgment" can be applied directly and the proportion of times A exceeds B can be translated directly into a measure of standardized mean difference between A and B (see Torgerson, 1958, p. 159ff).

We have applied this procedure in connection with a meta-analysis of research on the relationship of class-size to achievement (Glass and Smith, 1979).

Only the post-1960 studies were included in the scaling analysis. The regression analyses show that studies done prior to 1960 showed little relationship between class-size and achievement (probably because of poor design, poor measures, and because genuinely small classes--less than a dozen pupils, say--were seldom studied). The post 1960 studies produced 246 values of Δ_{S-L}, for which one needs only to note whether Δ is positive or negative. In addition, there were a small number of studies that yielded only comparisons of the sizes of the achievement means for the small and large classes, but no metric information from which Δ might be calculated. The principal study of this type was Forno and Collins (1967). The findings from these studies could be included in the scaling analyses even though they could not be included in the regression analyses. The total number of paired comparisons was 559.

The class-size dimension was broken into five categories in an attempt to obtain an even distribution of comparisons. These categories were as follows: 1-11 pupils, 12-22, 23-32, 33-42, 43 or more pupils. The actual average class-sizes falling into these categories were as follows: 2, 18, 28, 38, and 84 pupils. These averages will be used to represent the categories. Thus, a comparison of achievement means for classes of sizes 4 and 30, for example, will be spoken of as a comparison of classes of size 2 and 28.

The following frequency matrix was obtained by counting direction of superiority in the paired comparisons:

Paired Comparison Frequency Matrix

	Class Size				
	2	18	28	38	84
2	-	7 of 8	45 of 46	3 of 3	
18	1 of 8	-	111 of 160	124 of 157	2 of 3
28	1 of 46	49 of 160	-	109 of 167	5 of 9
38	0 of 3	33 of 157	58 or 167	-	1 of 6
84		1 of 3	4 of 9	5 of 6	-

This matrix is read as follows: each entry represents the number of times the row class-size had a higher achievement mean than the column class-size. For example, there were 46 comparisons of class-size 2 and class-size 28; in 45 of them, achievement was superior in the class of 2.

It was decided at this point that some comparisons were so infrequently represented that including them in the scaling analysis might greatly overweight their unstable estimates. It was decided arbitrarily to include only those cells with more than a half-dozen comparisons. Thus, the following three cells (three on each side of the diagonal) were eliminated: row 1 - column 4; row 2 - column 5; row 4 - column 5. The resulting frequency matrix is then transformed to a proportions matrix, π, e.g., 111 of 160 = .69 and then to an $\underline{X}$-matrix where X_{ij} is the unit normal deviate below which lies π_{ij} proportion of the normal curve. The π and $\underline{X}$ matrices are combined in the following figure:

	2	18	28	38	84
2	-	π = .88 X = 1.18	.98 2.05		
18	.12 -1.18	-	.69 .50	.79 .81	
28	.02 -2.05	.31 -.50	-	.65 .39	.56 .15
38		.21 -.81	.35 -.39	-	
84			.44 -.15		-

The solution for scale values follows Gulliksen's (1956) least-squares solution for incomplete data. A vector Z is formed by summing the columns of $\underline{X}$: $-Z^T = (3.23, 0.57, -2.01, -1.20, -0.15)$. A matrix of $\underline{M}$ of order 5x5 is formed such that a -1 is entered in each off-diagonal cell in $\underline{X}$ that is not empty, a zero is entered for each empty cell, and the diagonal entry is the number of non-empty cells in the corresponding column of X. The last scale value, corresponding to class-size 84, is arbitrarily set equal to zero, and the last row and column of M are deleted. The reduced matrices, $\underline{M}_1$ and $\underline{Z}_1$, are combined to form the normal equations of the least-squares solution for $\underline{S}_1$ the scale values:

$$S_1 = M_1^{-1} Z_1$$

The estimates and their solution are as follows:

$$S_1 = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 3.23 \\ 0.57 \\ -2.01 \\ -1.20 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 1.625 & 1.250 & 1.000 & 1.125 \\ 1.250 & 1.500 & 1.000 & 1.250 \\ 1.000 & 1.000 & 1.000 & 1.000 \\ 1.125 & 1.250 & 1.000 & 1.625 \end{bmatrix} \begin{bmatrix} 3.23 \\ 0.57 \\ -2.01 \\ -1.20 \end{bmatrix}$$

$$S^T = (2.60,\ 1.38,\ 0.59,\ 0.39,\ 0)$$

The graph of the scaled relationship between class-size and achievement appears as Figure 9. The scale values on the ordinate of the graph are arbitrary. The quadratic equation which best fits the five points by the least-squares criterion is as follows:

$$s = 2.78912 - 0.09318(\text{Size}) + 0.000715(\text{Size})^2$$

The multiple R-squared is 0.99. The following estimates of achievement (on an arbitrary scale) for various class-sizes were obtained from the regression curve:

Size	Estimated Scale Value for Achievement	Decrease in Achievement From 10 More Pupils
1	2.70	.86
10	1.93	.72
20	1.21	.57
30	0.64	.43
40	0.21	.29
50	-0.08	.15
60	-0.23	0
70	-0.23	
80	-0.09	

The curve in Figure 9 shows the expected and quite plausible decreasing deceleration in achievement as class-size increases.

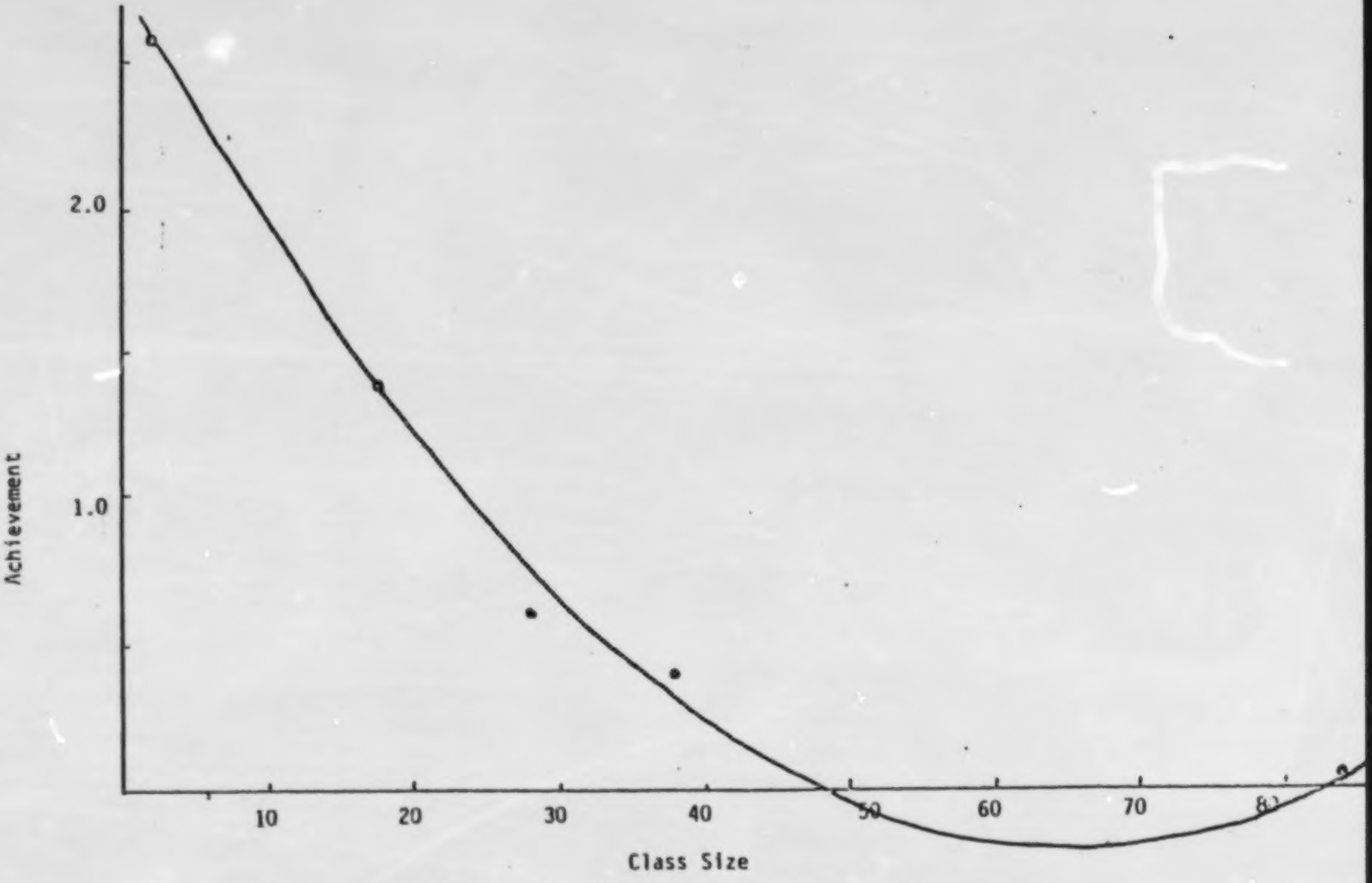


Figure 5.1 Relationship between class-size and achievement (arbitrary units) obtained by psychometric scaling of comparisons

FINDINGS OF EXPERIMENTAL STUDIES

The description of findings in experimental studies so that results can be aggregated and their variability studies present several technical problems. The findings of comparative experiments are probably best expressed as standardized mean differences between pairs of treatment conditions. It will seldom be satisfactory to express experimental findings as a measure of association between several levels of an independent variable and a metric dependent variable. Such association measures (e.g., ω^2) are descriptive of a complete, somewhat arbitrary, set of experimental conditions an investigator chooses to investigate in a single study. For example, if one wished to determine the comparative effects of computer-assisted and traditional foreign language instruction, then it is irrelevant that a televised instruction condition was also present in a study, and one would not want a quantitative measure of effect to be influenced by the irrelevant condition (Glass & Hakstian, 1969).

In what follows, reference will be made to the comparison of a particular experimental condition with a control group. Of course, there may be no "control" group in a traditional sense, and one could imagine that two different experimental conditions are compared. The most informative and straightforward measure of experimental effect size is the mean difference divided by within-group standard deviation:

$$\Delta = \frac{\bar{X}_E - \bar{X}_C}{S_x} \qquad (1)$$

Suppose that four experiments were performed in which either nialonide or iproniazid was compared with a placebo for efficacy in relieving depression. Three of the experiments measured outcomes with the MMPI D scale; the fourth study used the Beck Depression Inventory. Suppose the following results were obtained. (The data are hypothetical, but the findings are close to those reported in Smith, Glass and Miller (1980)).

Study No.	Comparison	Test	Means	St. dev.	A-B
1	Nialomide vs. Placebo	MMPI	70.10-70.50	9.50	-.04
2	Nialomide vs. Placebo	MMPI	61.45-62.31	11.25	-.08
3	Iproniazid vs. Placebo	MMPI	60.21-65.15	7.80	-.63
4	Iproniazid vs. Placebo	Beck	110.75-121.45	20.50	-.52

In the above data, the average effect of nialomide is -.06, i.e., six-hundreths standard deviation superior to a placebo; the average effect of iproniazid is -.58, more than a half standard deviation.

The meaning of Δ is readily comprehended and, assuming some distribution form, can be translated into notions of overlapping distributions of scores and comparable percentiles. For example, suppose that a study of the effect of ritalin versus placebo on reducing hyperactivity reveals an Δ of -1.00. One knows immediately that the average child on ritalin shows hyperactivity one standard deviation below that of the average child on placebo; thus, assuming normality, only 16 percent of the placebo children are less hyperactive than the average child on the drug, and so on.

Another way to interpret the magnitude of the effect size is to compare it to other effect sizes, particularly for effects that many people have external references for how strong the treatment was. One TV program that the American public has enthusiastically endorsed is Sesame Street. Effects of Sesame Street on social behavior, such as cooperation, were included in a meta-analysis. However, the primary aim of Sesame Street, particularly the first year, was cognitive skills instruction -- prereading, language, and math. These cognitive outcome measures were not considered in the meta-analysis, but are considered by many parents and preschool teachers to be substantial.

In 1970 and again in 1971, the Educational Testing Service (ETS) conducted a field study evaluation of Sesame Street. Both years had numerous measurements, several subsamples, several research designs, and confounded results making a single numerical summary statement difficult. The most easily interpreted results compared two groups of 4½ to 5 year old disadvantaged children, of which one group had not seen Sesame Street while the other had watched for one season. The criterion measure was a special test developed by ETS covering the cognitive skills taught on the program. The tendency was for those who watched more to gain more although viewing differences were confounded with intelligence and other background variables. The effect sizes for four levels of viewing versus no viewing all favor the Sesame Street viewers, varying from .53 to 1.45, with a mean of 1.00.

A more controlled analysis was possible the second year with 283 children randomly assigned to groups who either had or did not have a TV to view the program. A set of covariance analyses (covarying on pretest score, pretest Peabody IQ, and SES) resulted in seven effect sizes, varying from .04 to .54. Dropping the "parts of whole" test that was a low outlier the mean effect size was .45 with a standard deviation of .085. The remaining tests covered the topics of number, sorting, forms, pre-reading, relational terms, and classification.

Electric Company, the Sesame Street sequel for older children, was evaluated by ETS in 1973 and 1974. Again, there were numerous analyses, but using the total score on an ETS reading test as the criterion measure, for children in grades one to four in two cities comparing those who were encouraged to watch the program at home versus those who were not encouraged the average effect size was .17. This effect is low partially because non-encouraged children also watched the program, thus this effect size is a measure of increased reading achievement due to increased watching when encouraged by a teacher to view the program after school.

Both the first and second year evaluation also had an in-school experimental design component. Two locations with large numbers of either Spanish speaking or black children were assigned to teachers who were encouraged to show the program regularly during the year or who were asked not to. The amount of viewing and supplemental instruction was teacher determined. Two outcome measures, the ETS reading test and the Metropolitan Achievement Test provided similar results. Averaging the data from two locations, grades one through three and the two years, resulted in an effect size of .43 (S.D. = .30) for the ETS reading test and .35 (S.D. = .31) for the Metropolitan Achievement test. The overall average is .39, with scores ranging from -.03 to 1.02.

Interpretations of effect sizes, Δ_{E-C}, in terms of percentiles (e.g., if $\Delta = +1.00$, then the average person in the experimental group has a score that exceeds 84 percent of the persons' scores in the control group) depend, of course, on assumptions about the shapes of the distributions of the variable in the two groups. Normality is a convenient and unobjectionable assumption in many instances, but its convenience should not blind one to the fact that it is an assumption that may occasionally be false. Kraemer and Andrews (1980) have called attention to this problem. Suppose, for example, that the scores in the experimental and control groups are distributed according to the exponential distribution (Hastings and Peacock, 1974, pp. 56-59) with the following parameters:

Group	Distribution	Mean	St. dev.
Experimental	$P(X_E) = a_1 e^{-a_1 x}$	$1/a_1$	$1/a_1$
Control	$P(X_c) = a_2 e^{-a_2 x}$	$1/a_2$	$1/a_2$

Now the effect size Δ_{E-C} equal to

$$\frac{\bar{X}_E - \bar{X}_C}{s_c}$$

will estimate, in the case of exponential distributions,

$$\Delta = \frac{1/a_1 - 1/a_2}{1/a_2}$$

$$= \frac{a_2 - a_1}{a_1} \qquad (2)$$

Suppose that a particular experiment yields summary statistics as follows:

$$\bar{X}_E = 18 \quad , \; s_E = 16 \; ;$$

$$\bar{X}_C = 10 \quad , \; s_C = 10 \; .$$

The value of Δ equals (18 - 10) / 8 = +1. If it is assumed that the two distributions are normal, then the Δ of +1 has the usual interpretation: the average person in the experimental group exceeds 84 percent of the persons in the control group. Suppose, however, that the average experimental group person's score is expressed as a percentile in the control group, assuming exponential distribution in each group. Then the percentile rank of X = 18 in an exponential distribution with paramenter $a_2 = 10$ is given by

$$\int_0^{18} P(x)dx = \int_0^{18} 10\, e^{-10X}\, d_X = .834 \quad .$$

Thus, assuming exponential distributions within essentially experimental and control groups gives essentially the same interpretation of +1 as the assumption of normal distrubutions (.83 vs. .84). This example is not meant to suggest that the exponential distribution is in any sense interchangeable as an assumption with the normal distribution. The assumption of distribution shapes may be important and it should be checked when possible and the most reasonable assumption made.

The choice of the standard deviation with which to scale the differences between group means to determine Δ is crucial. Various choices can result in substantial differences in effect size.

The definition of Δ appears uncomplicated, but heterogeneous group variances cause difficulties. Suppose that experimental and control groups have means and standard deviations as follows:

	Experimental	Control
Means	$\bar{y}_E = 52$	$\bar{y}_C = 50$
Standard Deviations	$S_E = 2$	$S_C = 10$

The measure of experimental effect could be calculated either by use of S_E or S_C or some combination of the two.

Basis of Standardization	Δ
a) S_E	1.00
b) S_C	0.20
c) $(S_E + S_C)/2$	0.33

The average standard deviation, c), probably should be eliminated as a mere mindless statistical reaction to a perplexing choice. But both the remaining 1.00 and 0.20 are correct; neither can be ruled out as false. It is true, in fact, that the experimental group mean is one standard deviation above the control group mean in terms of the experimental group standard deviation; and, assuming normality, the average subject in the control group is superior to only 16 percent of the members of the experimental group. However, the control group mean is only one-fifth standard deviation below the mean of the experimental group when measured in control group standard deviations; thus, the average experimental group subject exceeds 58 percent of the subjects in the control group. These facts are not contradictory; they are two distinct features of a finding which cannot be expressed by one number. In a meta-analysis of psychotherapy experiments, the problem of heterogeneous standard deviations was resolved from a quite different direction. Suppose that methods A, B, and Control are compared in a single experiment, with the following results:

	Method A	Method B	Control
Means	50	50	48
Standard deviations	10	1	4

If effect sizes are calculated using the standard deviations of the "method," then Δ_A equals 0.20 and Δ_B equals 2.00 -- a misleading difference, considering the equality of the method means on the dependent variable. Standardization of mean differences by the control group standard deviation at least has the advantage of allotting equal effect sizes to equal means. This seems reason enough to resolve the choice in favor of the control group standard deviation, at least when there are more than two treatment conditions and only one control condition.

Estimation of Δ

Given that

$$\Delta_{A-B} = \frac{\mu_A - \mu_B}{\sigma_y} , \qquad (3)$$

and assuming for the moment an understanding of which of many possible choices of σ_y is implied, the intuitively reasonable estimator of Δ is

$$\hat{\Delta}_{A-B} = \frac{\bar{y}_A - \bar{y}_B}{s_y}, \tag{4}$$

where the sample means are conventionally defined and s_y is the square root of the unbiased estimator of σ_y^2. Hedges (1979) showed the error of intuition with regard to (4), and he derived the maximum likelihood estimator of Δ assuming normality and a single sample estimate of σ_y.

Hedges (1979) examined the statistical properties of

$$\hat{\Delta}_{E-C} = \frac{\bar{X}_E - \bar{X}_C}{s_C}$$

as an estimator of

$$\Delta_{E-C} = \frac{\mu_E - \mu_C}{\sigma C}$$

He was able to show that

$\hat{\Delta}_{E-C} (n_1 n_2/(n_1 + n_2))^{\frac{1}{2}}$ is distributed

as a non-central t variate with non-centrality parameter

$\Delta_{E-C} (n_1 n_2/(n_1 + n_2))^{\frac{1}{2}}$ and degrees of freedom equal

to $n_2 - 1$ where n_1 and n_2 are the sizes of the samples for the experimental and control groups, respectively. Of course, this finding rests on the assumption that X is normally distributed for both the experimental and control groups.

It followed as a consequence of this theorem that the expected value of $\hat{\Delta}_{E-C}$ is given by

$$E(\hat{\Delta}) = \Delta . \left[K(n_2 - 1)\right]^{-1}, \text{ where}$$

$$K(n_2-1) = \frac{\Gamma\left(\frac{n_2-1}{2}\right)}{\sqrt{\frac{n_2-1}{2}}\;\Gamma\left(\frac{n_2-2}{2}\right)} . \qquad (5)$$

Hence, $\hat{\Delta}$ is biased as an estimator of Δ . The degree of bias is a function of the ratio of two gamma distributions as can be seen above. In Figure 5.2 (from Hedges, 1979), the bias in $\hat{\Delta}$ as an estimator of Δ is depicted by graphing the ratio $E(\hat{\Delta}) / \Delta$ against n_2-1. As can be seen there, $\hat{\Delta}$ is positively biased for small n; beyond sample size n_2 of 20, the bias is 10 percent of less.

Clearly, an unbiased estimate of Δ could be obtained by multiplying $\hat{\Delta}$ by the correction factor $K(n_2-1)$. Hedges (1979, p. 11) provided a table of values of $K(n_2-1)$ which is reproduced as Table 5.3,slightly modified form with his kind permission.

Hedges (1979) pointed out an unexpected and important property of effect sizes as estimators. Suppose that one obtains a series of observations of effect sizes, $\hat{\Delta}_i$, each of which estimates the same parameter value Δ . Assume further that for J such estimates, an aggregate estimate is obtained by averaging; thus

$$\Delta \text{ is estimated by } \sum_{I}^{J} \hat{\Delta}_i / J.$$

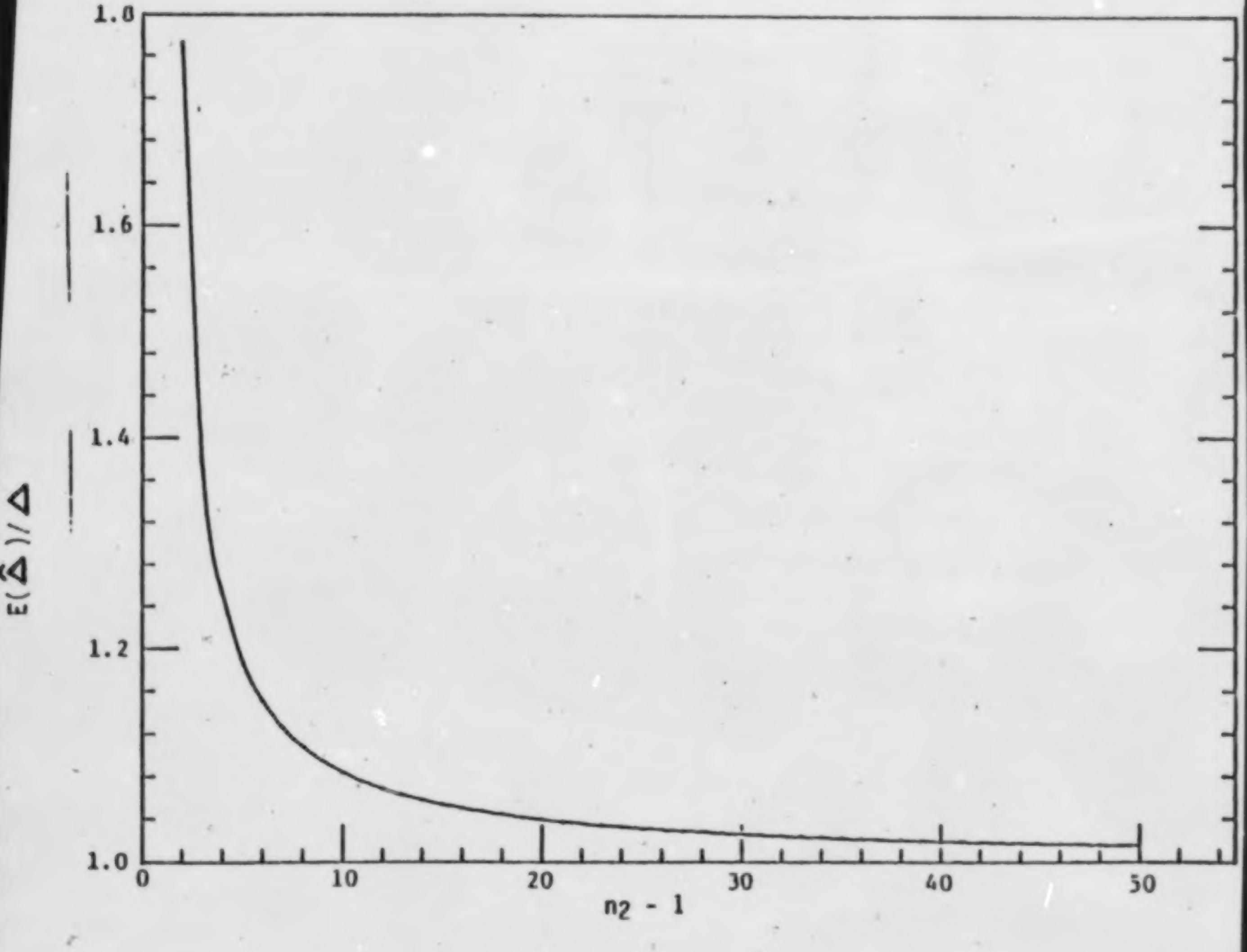


Figure 5.2. Ratio of the expected value of the estimated effect size to the parameter value as a function of the control group sample size, n_2.

Table 5.3

Value of $K(n_2-1)$ for n_2 to be used in obtaining unbiased estimates of Δ

n_2-1	K	n_2-1	K	n_2-1	K
2	0.56419	21	0.96378	40	0.98111
3	0.72360	22	0.96545	41	0.98158
4	0.79788	23	0.96697	42	0.98202
5	0.84075	24	0.96837	43	0.98244
6	0.86863	25	0.96965	44	0.98284
7	0.88820	26	0.97083	45	0.98322
8	0.90270	27	0.97192	46	0.98359
9	0.91387	28	0.97293	47	0.98394
10	0.92275	29	0.97387	48	0.98428
11	0.92996	30	0.97475	49	0.98460
12	0.93594	31	0.97558	50	0.98491
13	0.94098	32	0.97635		
14	0.94529	33	0.97707		
15	0.94901	34	0.97775		
16	0.95225	35	0.97839		
17	0.95511	36	0.97900		
18	0.95765	37	0.97957		
19	0.95991	38	0.98011		
20	0.96194	39	0.98062		

Denote this latter estimator by *G*, as did Hedges. He showed that ". . . G is *not* a consistent estimator of Δ as $J \rightarrow \infty$. That is, even though the number of experiments combined increases, the estimator does not necessarily approximate the true value Δ more closely. In fact, the estimates can differ from Δ by a considerable amount depending on the sample sizes. To see this, consider the example of a collection of experiments with 5 subjects per group. The estimator $\hat{\Delta}$ has a bias which results in overestimation of Δ by approximately 25 percent when four degrees of freedom are used for σ. Each estimator $\hat{\Delta}_i$ has the same bias, therefore G is biased by the same amount as each $\hat{\Delta}_i$, $i = 1, \ldots, J$. As J increases, the bias is unchanged, but the variance of G tends to zero. Thus as the number of studies increases, the estimator G estimates the wrong quantity more precisely."

(Hedges, 1979, pp. 8-9;
notation altered slightly.)

The inconsistency in G as an estimator of Δ can be corrected by using Hedges' earlier result, viz., correct each estimate $\hat{\Delta}_i$ by $\underline{K}\,(\underline{n}_2-1)$ before averaging them.

Although Δ is simple, it can present many difficulties in both conception and execution. Many research reports do not contain the means and standard deviations of experimental conditions. Where there are more than two experimental conditions and means are not reported, there is little hope of ever recovering an Δ from the report. There are several circumstances of incomplete data reporting in which a harmless assumption and some simple algebra will make it possible to reconstruct Δ measures:

1. One knows the value of t and whether $\overline{X}_E$ or $\overline{X}_C$ is larger.
2. One knows the significance level of a mean difference and the two sample sizes.
3. One knows $\overline{X}_{E_1}$, $\overline{X}_{E_2}$, . . ., and the value of F.
4. One knows $\overline{X}_E$ and $\overline{X}_C$ and the value of some multiple comparisons statistics such as Tukey's q or Dunn's or Dunnett's statistics.

One example worked out in detail should suffice to illustrate how to proceed in these general circumstances. The report of an experiment contains J means $\overline{X}_1$, $\overline{X}_2$, . . . ,$\overline{X}_J$, the sizes of each group (n_1, . . . , n_J), and an F statistic. Suppose that $\overline{X}_1$ is the mean of the experimental condition of interest and that a second condition is a control yielding $\overline{X}_c$.

The value of the F statistic was calculated by the original investigator from the following formula:

$$F=\frac{\Sigma n_j(\overline{X}_j-\overline{X})^2/(J-1)}{\Sigma(n_j-1)s^2_j/(N-J)}=\frac{MS_B}{MS_w},$$

where the only symbol which might not be obvious is N, which equals $n_1 + n_2 + \ldots + n_J$. Under the assumption that the variance S_J^2, in each group is the same, the above expression can be readily solved to obtain S_x^2, the assumed homogeneous variance:

$$s_x^2=\frac{MS_B}{F}.$$

The effect size follows directly:

$$\Delta=\frac{\overline{X}_1-\overline{X}_C}{s_x}.$$

How to calculate Δ when S_j^2 is not homogeneous and how to define S_X in multifactor experimental designs are more than simple technical questions. As will be seen later in this chapter, they raise basic concerns about the definition and meaning of Δ .

One commonly encountered method of reporting results presents unique difficulties. Reports sometimes give only the sample sizes and an indication of whether a mean difference was statistically significant at a customary level. A conservative approximation to the Δ can be derived by setting a t-ratio equal to the critical value corresponding to the reported significance level and solving for $(\overline{X}_E - \overline{X}_C)/ S_X$, under the assumption of equal within-group variances. For example, suppose that a report contains only the information that the mean of the $\underline{n}_1$ experimental subjects exceeded the mean of the $\underline{n}_2$ control subjects at the .05 level of significance. At the very least, then,

$$t = \frac{\overline{X}_E - \overline{X}_C}{\sqrt{s_x^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 1.96.$$

Clearly,

$$\Delta = \frac{\overline{X}_E - \overline{X}_C}{s_x} = 1.96\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

gives a conservative estimate of the experimental effect. This small bit of algebra also indicates how one obtains Δ when given only t and $\underline{n}_1$ and $\underline{n}_2$:

$$\Delta = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \qquad (6)$$

When the n's in the two groups are equal, the effect size is simply the value of the t-statistic multiplied by the square root of the ratio of 2 to n, the common sample size. This calculation permits a two-way tabulation in which Δ can be found given t and n. Such a table is reproduced as Table 5.4. As an illustration of how it is read, consider a study in which the means of two groups of 12 persons each were compared with a t-test and a t-statistic of +2.10 was obtained. From the table, the value of Δ is +.86.

The Homogeneity of Variances Assumption in Transforming t and F Statistics

In many studies where the emphasis in reporting is on inferential statistics, only pooled information is available about the within-group variances. Since the statistical tests used in these cases depend on an assumption of homogeneity of within-group variances, the test statistics frequently obscure whatever differences in variance might have existed.

When the results of an experiment are expressed as a t-statistic which is reported along with n_1, and n_2 but without means and variances, one can calculate an effect-size, Δ, via the formula

$$\Delta_p = t(1/n_1 + 1/n_2)^{\frac{1}{2}} . \tag{7}$$

The subscript p indicates that Δ is based on a "pooling" of variances. Suppose, to the contrary, that the sample variances are unequal, and that one wishes Δ_C, the mean difference standardized by the control group (group 1, for example) standard deviation.

Table 5.4. Table for converting t-statistic to effect size, Δ given equal sample sizes, n.

n

t	6	8	10	12	14	16	18	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.10	.06	.05	.04	.04	.04	.04	.03	.03	.03	.03	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.01	.01	.01
.20	.12	.10	.09	.08	.08	.07	.07	.06	.06	.05	.05	.04	.04	.04	.04	.04	.04	.03	.03	.03	.03	.03	.03	.03
.30	.17	.15	.13	.12	.11	.11	.10	.09	.08	.08	.07	.07	.06	.06	.06	.05	.05	.05	.05	.05	.05	.04	.04	.04
.40	.23	.20	.18	.16	.15	.14	.13	.13	.11	.10	.10	.09	.08	.08	.08	.07	.07	.07	.07	.06	.06	.06	.06	.06
.50	.29	.25	.22	.20	.19	.18	.17	.16	.14	.13	.12	.11	.11	.10	.10	.09	.09	.08	.08	.08	.08	.07	.07	.07
.60	.35	.30	.27	.24	.23	.21	.20	.19	.17	.15	.14	.13	.13	.12	.11	.11	.11	.10	.10	.09	.09	.09	.09	.08
.70	.40	.35	.31	.29	.26	.25	.23	.22	.20	.18	.17	.16	.15	.14	.13	.13	.12	.12	.11	.11	.11	.10	.10	.10
.80	.46	.40	.36	.33	.30	.28	.27	.25	.23	.21	.19	.18	.17	.16	.15	.15	.14	.14	.13	.13	.12	.12	.12	.11
.90	.52	.45	.40	.37	.34	.32	.30	.28	.25	.23	.22	.20	.19	.18	.17	.16	.16	.15	.15	.14	.14	.13	.13	.13
1.00	.58	.50	.45	.41	.38	.35	.33	.32	.28	.26	.24	.22	.21	.20	.19	.18	.18	.17	.16	.16	.15	.15	.15	.14
1.10	.64	.55	.49	.45	.42	.39	.37	.35	.31	.28	.26	.25	.23	.22	.21	.20	.19	.19	.18	.17	.17	.16	.16	.16
1.20	.69	.60	.54	.49	.45	.42	.40	.38	.34	.31	.29	.27	.25	.24	.23	.22	.21	.20	.20	.19	.18	.18	.17	.17
1.30	.75	.65	.58	.53	.49	.46	.43	.41	.37	.34	.31	.29	.27	.26	.25	.24	.23	.22	.21	.21	.20	.19	.19	.18
1.40	.81	.70	.63	.57	.53	.49	.47	.44	.40	.36	.33	.31	.30	.28	.27	.26	.25	.24	.23	.22	.21	.21	.20	.20
1.50	.87	.75	.67	.61	.57	.53	.50	.47	.42	.39	.36	.34	.32	.30	.29	.27	.26	.25	.24	.24	.23	.22	.22	.21
1.60	.92	.80	.72	.65	.60	.57	.53	.51	.45	.41	.38	.36	.34	.32	.31	.29	.28	.27	.26	.25	.25	.24	.23	.23
1.70	.98	.85	.76	.69	.64	.60	.57	.54	.48	.44	.41	.38	.36	.34	.32	.31	.30	.29	.28	.27	.26	.25	.25	.24
1.80	1.04	.90	.80	.73	.68	.64	.60	.57	.51	.46	.43	.40	.38	.36	.34	.33	.32	.30	.29	.28	.28	.27	.26	.25
1.90	1.10	.95	.85	.78	.72	.67	.63	.60	.54	.49	.45	.42	.40	.38	.36	.35	.33	.32	.31	.30	.29	.28	.28	.27
2.00	1.15	1.00	.89	.82	.76	.71	.67	.63	.57	.52	.48	.45	.42	.40	.38	.37	.35	.34	.33	.32	.31	.30	.29	.28
2.10	1.21	1.05	.94	.86	.79	.74	.70	.66	.59	.54	.50	.47	.44	.42	.40	.38	.37	.35	.34	.33	.32	.31	.30	.30
2.20	1.27	1.10	.98	.90	.83	.78	.73	.70	.62	.57	.53	.49	.46	.44	.42	.40	.39	.37	.36	.35	.34	.33	.32	.31
2.30	1.33	1.15	1.03	.94	.87	.81	.77	.73	.65	.59	.55	.51	.48	.46	.44	.42	.40	.39	.38	.36	.35	.34	.33	.33
2.40	1.39	1.20	1.07	.98	.91	.85	.80	.76	.68	.62	.57	.54	.51	.48	.46	.44	.42	.41	.39	.38	.37	.36	.35	.34
2.50	1.44	1.25	1.12	1.02	.94	.88	.83	.79	.71	.65	.60	.56	.53	.50	.48	.46	.44	.42	.41	.40	.38	.37	.36	.35
2.60	1.50	1.30	1.16	1.06	.98	.92	.87	.82	.74	.67	.62	.58	.55	.52	.50	.47	.46	.44	.42	.41	.40	.39	.38	.37
2.70	1.56	1.35	1.21	1.10	1.02	.95	.90	.85	.76	.70	.65	.60	.57	.54	.51	.49	.47	.46	.44	.43	.41	.40	.39	.38
2.80	1.62	1.40	1.25	1.14	1.06	.99	.93	.89	.79	.72	.67	.63	.59	.56	.53	.51	.49	.47	.46	.44	.43	.42	.41	.40
2.90	1.67	1.45	1.30	1.18	1.10	1.03	.97	.92	.82	.75	.69	.65	.61	.58	.55	.53	.51	.49	.47	.46	.44	.43	.42	.41
3.00	1.73	1.50	1.34	1.22	1.13	1.06	1.00	.95	.85	.77	.72	.67	.63	.60	.57	.55	.53	.51	.49	.47	.46	.45	.44	.42
3.10	1.79	1.55	1.39	1.27	1.17	1.10	1.03	.98	.88	.80	.74	.69	.65	.62	.59	.57	.54	.52	.51	.49	.48	.46	.45	.44
3.20	1.85	1.60	1.43	1.31	1.21	1.13	1.07	1.01	.91	.83	.76	.72	.67	.64	.61	.58	.56	.54	.52	.51	.49	.48	.46	.45
3.30	1.91	1.65	1.48	1.35	1.25	1.17	1.10	1.04	.93	.85	.79	.74	.70	.66	.63	.60	.58	.56	.54	.52	.51	.49	.48	.47
3.40	1.96	1.70	1.52	1.39	1.29	1.20	1.13	1.08	.96	.88	.81	.76	.72	.68	.65	.62	.60	.57	.56	.54	.52	.51	.49	.48
3.50	2.02	1.75	1.57	1.43	1.32	1.24	1.17	1.11	.99	.90	.84	.78	.74	.70	.67	.64	.61	.59	.57	.55	.54	.52	.51	.49
3.60	2.08	1.80	1.61	1.47	1.36	1.27	1.20	1.14	1.02	.93	.86	.80	.76	.72	.69	.66	.63	.61	.59	.57	.55	.54	.52	.51
3.70	2.14	1.85	1.65	1.51	1.40	1.31	1.23	1.17	1.05	.96	.88	.83	.78	.74	.71	.68	.65	.63	.60	.59	.57	.55	.54	.52
3.80	2.19	1.90	1.70	1.55	1.44	1.34	1.27	1.20	1.07	.98	.91	.85	.80	.76	.72	.69	.67	.64	.62	.60	.58	.57	.55	.54
3.90	2.25	1.95	1.74	1.59	1.47	1.38	1.30	1.23	1.10	1.01	.93	.87	.82	.78	.74	.71	.68	.66	.64	.62	.60	.58	.57	.55
4.00	2.31	2.00	1.79	1.63	1.51	1.41	1.33	1.26	1.13	1.03	.96	.89	.84	.80	.76	.73	.70	.68	.65	.63	.61	.60	.58	[illegible]

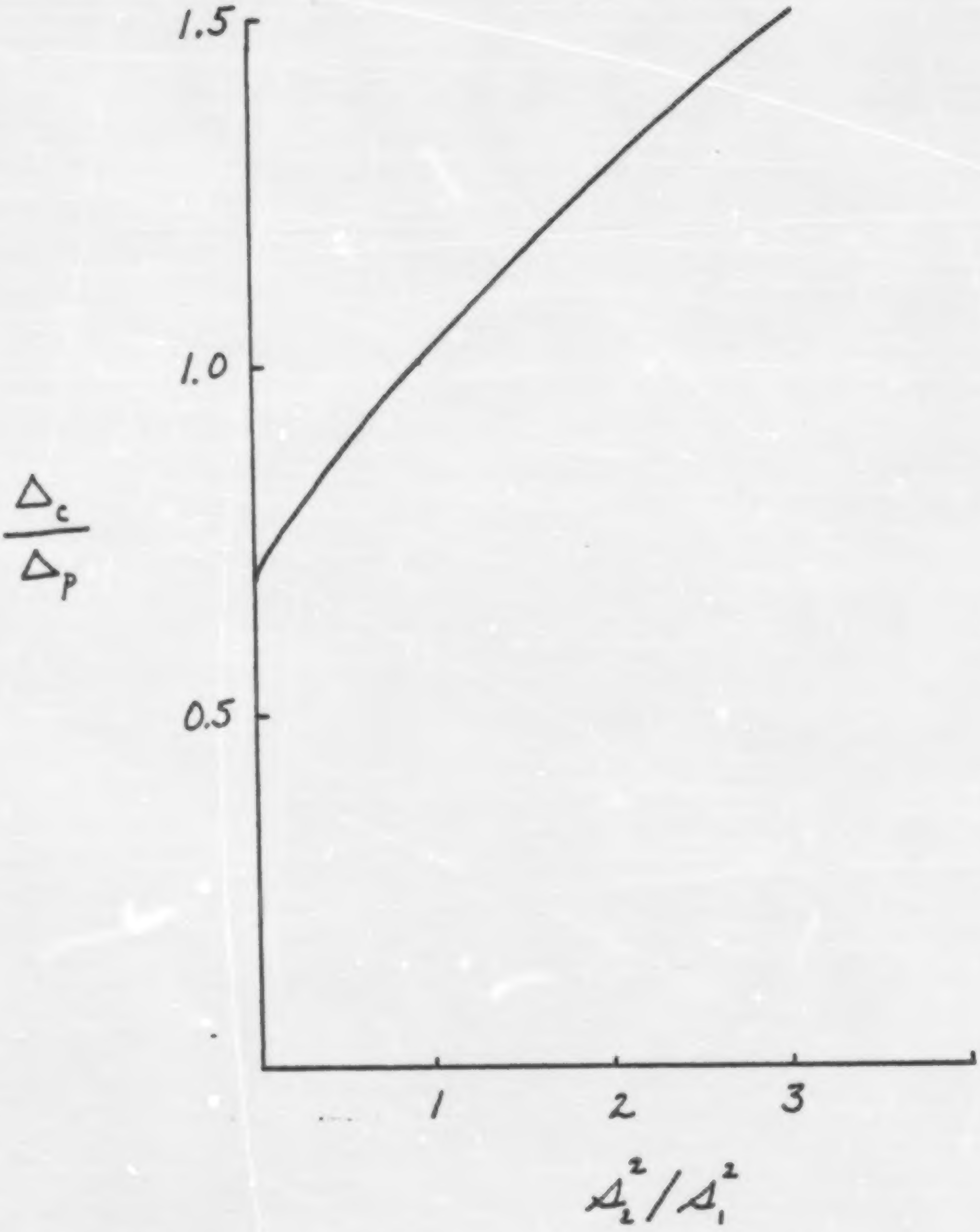


Table 5.3. Relationship of control vs. experimental group variance ratio to bias in effect-size approximation.

Assuming $\underline{n}_1 = \underline{n}_2$, the ratio of Δ_C to Δ_p can be derived:

$$\frac{\Delta_C}{\Delta_p} = \left[\left(1 + \frac{s_2^2}{s_1^2}\right) \div 2 \right]^{\frac{1}{2}} \qquad (8)$$

As can be seen in Formula (8) $\underline{\Delta}_C$ is exactly equal to $\underline{\Delta}_p$ when variances are equal. The bias in the approximation is negative and no greater than about 25 percent when control group variance is less than experimental group variance; however, the bias can grow beyond any bounds when the inequality in the variances is reversed.

As can be seen in Figure 5.3, $\underline{\Delta}_C$ is exactly equal to the surrogate, but accesible, value $\underline{\Delta}_p$ when variances are equal. The bias in the approximation is negative and no greater than about 25 percent when control group variance is less than experimental group variance; however, the bias can grow beyond any bounds when the inequality in the variances is reversed. This indicates to us that the approximation of $\underline{\Delta}_C$ via a $\underline{t}$-statistic (or presumably an $\underline{F}$-ratio, as well) could be unsafe if the sample variance of the experimental group substantially exceeds that for the control group.

A psychological experiment performed by Hekmat (1973) illustrates the problems of this section and concerns of earlier sections about choice of the control group standard deviation and non-normality. Hekmat compared three methods of treating a phobia against an untreated control group. Ten persons constituted each of the four groups. A

Table 5.5

Pre-Post Conditioning Means and Standard Deviations for the Behavioral Avoidance Test, Fear Survey Schedule

Group	Behavior avoidance test					Fear Survey Schedule			
	Pretest score			Post-conditioning score		Pretest score		Post-conditioning score	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Systematic desensitization	10	18.3	.82	5.0	3.39	4.6	.51	2.7	.67
Semantic desensitization	10	18.4	.87	4.5	3.17	4.6	.51	2.5	.84
Implosive therapy	10	18.0	.081	18.6	.96	4.6	.51	4.9	.31
Control	10	18.2	.78	17.8	.63	4.6	.51	4.4	.52

Note. On the Fear Survey Schedule, the pleasant pole scored 1, the unpleasant 7. The maximum phobia score was 5 which indicates "very much fear," and minimum score was 1 which indicates "no fear."

Behavior Avoidance Test and a Fear Survey Schedule were administered to each of the forty persons before and after the treatment. The means and standard deviations for the four groups on the two measures appear in Table 5.5.

Since persons were assigned randomly to groups, the pretest statistics may be disregarded. Notice the wide discrepancies among posttest standard deviations: on the BAT, the standard deviation for the systematic desensitization group is more than five times as great as that for the control group. If the effect size, Δ , comparing the systematic desensitization group against the control group is calculated by dividing by the experimental group standard deviation, its value is

$$\Delta = \frac{5.0 - 17.8}{3.39} = -3.78.$$

If, on the other hand, the control group standard deviation is used, the value of the effect size is

$$\Delta = \frac{5.0 - 17.8}{.63} = -20.32$$

An effect size of twenty standard deviations is an absurd figure.

Suppose that Hekmat had only reported t-statistics instead of means and standard deviations. The t-statistic for the comparison of the systematic desensitization and control groups would equal

$$t = \frac{5.0 - 17.8}{\sqrt{\frac{2}{10}\left(\frac{11.889}{2}\right)}} = -11.74$$

Converting this t-statistic to an effect size, assuming homogeneous variances as is necessary, gives a Δ of -5.25.

Effect sizes that bounce around from 20 to 3 to 5 to whatever else depending on one or another assumption indicate that something is fundamentally wrong. In the case of Hekmat's data the problem lies with the measurement scales. They undoubtedly would show, upon inspection of distributions of the data, severe ceiling and floor effects with resulting asymmetry and non-normality.

Studies Without Control Groups

Suppose that in a meta-analysis of experimental evaluations of science curricula that typical studies involve the comparison of a new curriculum (e.g., Science Curriculum Improvement Study (SCIS) or Science: A Process Approach (SAPA)) against traditional science curricula (group lecture, teacher-centered and oriented toward knowledge acquisition rather than developing inquiry skills). From such studies, effect sizes comparing SCIS or SAPA against Traditional could be calculated in the usual way, e.g.,

$$\Delta = \frac{\overline{Y}_{SAPA} - \overline{Y}_T}{s_T} ,$$

where the Traditional curriculum is thought of as a "control" condition.

Experiments will exist in which SCIS is compared to SAPA and no Traditional comparison is involved. It makes no sense to pool in the same analyses some effect sizes based on SCIS vs. Traditional comparisons, some based on SAPA vs. Traditional comparisons, and a third group based on SCIS vs. SAPA comparisons. For if SCIS and SAPA are both superior curricula, their large and positive effects should not be lumped with comparisons between themselves which would be small. The problem can be resolved by means of control referencing of the effect sizes. Each effect size based on a direct comparison of SCIS and SAPA can be broken into two effect sizes that reference the curriculum against a hypothetical control group (in this case, the Traditional curriculum).

Assume that there exists some number of effect sizes calculated from comparisons of SCIS and Traditional curricula; demote the average of these effects by $\bar{\Delta}_{SC}$. Likewise, denote the average of all effect sizes gotten by comparing SAPA and Traditional by $\bar{\Delta}_{SA}$. A single study in which SCIS and SAPA are compared without a Traditional group yields one effect size, Δ_{SC-SA}. We wish to break Δ_{SC-SA} into two effects, Δ'_{SC} and Δ'_{SA}, that estimate the effect sizes that would have been obtained in this study if a Traditional group had been included.

Two reasonable conditions may be imposed on Δ'_{SC} and Δ'_{SA}, the control-referenced effect sizes:

1) $\Delta_{SC-SA} = \Delta'_{SC} - \Delta'_{SA}$, and (9)

2) $\Delta'_{SC} - \bar{\Delta} = \bar{\Delta}_{SA} - \Delta'_{SA}$. (10)

These conditions imply 1) that the observed difference from the direct comparison is preserved in the control-referenced comparison, and 2) that the error (the deviation of a control-referenced effect from the average of all similar non-control-referenced effects) is equally shared between the two referenced effects. These two conditions establish a pair of independent linear equations in two unknowns that can be solved for the two control-referenced effects:

$$\Delta'_{SC} = (\Delta_{SC-SA} + \bar{\Delta}_{SC} + \bar{\Delta}_{SA})/2, \text{ and}$$

$$\Delta'_{SA} = \Delta'_{SC} - \Delta_{SC-SA}. \qquad (11)$$

Consider this illustration. In 100 comparisons of SCIS and Traditional curricula, the average effect size for the dependent variable "interest in science" is 0.76. For 200 comparisons of SAPA and Traditional, the average is 0.48. An experiment in which SCIS and SAPA were compared showed an effect size on "interest in science" of $\Delta_{SC-SA} = .30$. The two control-referenced effects, then, are given by

$$\Delta'_{SC} = (.30 + .76 + .48)/2 = .77, \text{ and}$$

$$\Delta'_{SA} = .77 - .30 = .47.$$

Finding a Standardizing Variance for Studies Without Control Groups.

Among the research reports relevant for a particular meta-analysis may be some which provide experimental comparisons of two treatment conditions of interest (say A and B) but include no control condition C. Such studies will provide, at best, standard deviations for the two treatment conditions but neither of these is appropriate for reasons discussed in the previous section. An estimate can be obtained however. If all studies in which A is compared with C are taken, the observed control group standard deviations can be regressed on the observed treatment A group standard deviations to give:

$$\hat{s}_C = b_0 + b_1 s_A \, . \qquad (12)$$

A similar regression can be established for s_C and s_B from those studies comparing treatment B with control C. Non-linear regressions are possible, of course. From a study comparing only treatments A and B, the observed standard deviations s_A and s_B can be substituted into their separate regression equations to provide two estimates of s_C. These two estimates could be pooled to provide the standard deviation with which to scale the mean difference $(\bar{y}_A - \bar{y}_B)$. From information from other studies about effect sizes for A and B against control, this effect between two treatment conditions could then be converted to separate effects between the treatment and control (see previous section). Experiments with quantitative independent variables (time, size, etc.) often have no untreated "control" condition. (A general approach to integrating effects from experiments with quantitative independent variables is described in Chapter Six.) For studies of drug dosage, amount of instruction and so on, a control condition of no treatment can be defined and included. For studies of an independent variable such as class size, one investigator's control can be another's treatment. But each study involves some number of comparisons of a small condition (S) and a large condition (L) and yields two means, $\bar{Y}_S$ and $\bar{Y}_L$, and two standard deviations s_S and s_L.

If the standard deviations vary with the value of the independent variable, then some value of that variable can be chosen as a reference point and its standard deviation used for converting all treatment mean differences to effect sizes. The problem is to find a way of converting from the observed s_S and s_L on the variable used in a given study to an estimate of s_R, the standard deviation for the reference group on that variable.

From all studies, the ratio of the observed standard deviations can be regressed on the values of the quantitative independent variable used in the comparison, viz., small (S) and large (L). The resulting regression function will be:

$$\left(\frac{s_S}{s_L}\right) = b_0 + b_1 S + b_2 L \ . \qquad (13)$$

If a standard deviation s_S is observed in a particular study for condition S, the standard deviation for the reference condition R could be estimated, if $R > S$, as:

$$\hat{s}_R = s_S/(b_0 + b_1 S + b_2 R) \ . \qquad (14)$$

A second estimate $\hat{s}_R$ can be obtained from the observed s_L in the same study. The mean of the two estimates could be used. (If $R < S$ or $R > L$, the regression equation can still be used but with substitutions appropriately reversed.) The observed mean differences $(\bar{y}_S - \bar{y}_L)$ can then be scaled to effect sizes for the corresponding differences in the value of the independent variable as:

$$\Delta_{S-L} = \frac{\bar{y}_S - \bar{y}_L}{\hat{s}_R} \ . \qquad (15)$$

Final Status Score

In a study with random assignment of subjects to treatment and control conditions, means can be obtained on a criterion measure Y as $\overline{Y}_T$ and $\overline{Y}_C$. The mean difference can be scaled to an effect size by the control group standard deviation on this measure, s_y. Final status, as the scale of the criterion measure, has several advantages over derived gain measures such as raw and residual gain scores and covariance adjusted final status scores. First, it is phenomenologically more relevant and, therefore, provides results more readily interpretable, particularly by lay audiences to whom a meta-analysis might be addressed. Second, the variance of the derived gain measures contain confounded "measurement error" which can significantly bias results.

Where there are pre-experiment group differences, the use of a post-treatment status scale will also be biased. It is with such biases that the derived gain measures were designed to deal. That they do not deal with them adequately is one problem. That they express the group comparisons on a scale different from that used in randomized studies with only a final status measure is a further problem for meta-analysis. If the final status scale is to be preferred then procedures must be found for converting results of studies using other scales to this one while minimizing the biases due to pre-experiment differences. This paper suggests such procedures.

Conversions From Other Scales

Raw Gain Scores. If the gain score from a pre-experiment measure (x) to a post-experiment criterion measure (y), for person i in the control group is:

$$G_{ci} = Y_{ci} - X_{ci} \qquad (16)$$

it is obvious that the mean gain is simply the difference between the post-experiment mean ($\bar{Y}_C$) and the pre-experiment mean ($\bar{X}_C$). The difference between treatment and control group means gains will be:

$$\bar{G}_T - \bar{G}_C = (\bar{Y}_T - \bar{Y}_C) - (\bar{X}_T - \bar{X}_C) \ . \qquad (17)$$

For the computation of an effect size on the final status scale the mean difference required is ($\bar{Y}_T - \bar{Y}_C$). It is better, however, to use ($\bar{G}_T - \bar{G}_C$). If there are no pre-treatment differences between the groups, i.e., ($\bar{X}_T - \bar{X}_C$) = 0, the two will be identical anyway. If there are pre-treatment differences, as there often are in studies in which gains are resorted to, then ($\bar{G}_T - \bar{G}_C$) has the advantage that it is not contaminated so directly by the pre-treatment differences.

Residual Scores. The residual element of the final status score, for person i in the control group, unexplainable from that person's status on a second variable X is:

$$\begin{aligned} g_{Ti} &= Y_{Ti} - \hat{Y}_{Ti} \\ &= Y_{Ti} - \{\bar{Y}_{..} + b_{y \cdot x}(X_{Ti} - \bar{X}_{..})\} \\ &= (Y_{Ti} - \bar{Y}_{..}) - b_{y \cdot x}(X_{Ti} - \bar{X}_{..}). \qquad (18) \end{aligned}$$

The mean difference between treatment and control groups in residual scores will be:

$$\bar{g}_T - \bar{g}_C = (\bar{Y}_T - \bar{Y}_C) - b_{y \cdot x}(\bar{X}_T - \bar{X}_C). \qquad (19)$$

Again, although the mean difference of interest for the computation of an effect size on the final status scale is ($\bar{Y}_T - \bar{Y}_C$), it is better to use ($\bar{g}_T - \bar{g}_C$). If there are no pre-experiment differences, the two will be the same. If there are, as there often are in studies in which residual scores are resorted to, then $(\bar{g}_T - \bar{g}_C)$ has the advantage that it is not contaminated so directly by the pre-treatment differences.

Covariance Adjusted Scores. Since the covariance adjustment of final scores is conceptually similar to the computation of residual final status scores, the same points may be made. The adjusted group means for ANCOVA will be the $\bar{g}$ in the previous section provided that the residuals there are computed using a regression line through the grand centroid $(\bar{X}_{..}, \bar{Y}_{..})$ with a pooled within-group estimate of slope.

Use of the regression line fitted to the total bivariate distribution, ignoring group membership, is inappropriate. If there is a treatment effect which shifts the relative levels group performance on Y, unpredictable from their relative positions on X, this treatment effect will be in part removed in the computation of the residuals. Use of a regression line through the grand centroid with a pooled within-group estimate of slope removes only those final status differences attributable to prior status differences and none due to treatment effects. If the total group regression line has been used in a study it will be difficult to include its results in a meta-analysis unless the prior and final status means are provided.

The difference between the covariance adjusted group means then, will be:

$$\begin{aligned}(\bar{Y}'_T - \bar{Y}'_C) &= (\bar{Y}_T - \bar{Y}_C) - b^{(w)}_{y\cdot x}\,(\bar{X}_T - \bar{X}_C) \\ &= \bar{g}_T - \bar{g}_C\,. \end{aligned} \qquad (20)$$

Achieving "Comparability" When There Are Pre-Treatment Group Differences. The uses of gain scores, residual scores and covairance adjustments when there are pre-experiment group differences are attempts to render the groups comparable. In a meta-analysis there is a different problem of comparability. If there are no pre-treatment differences, then mean differences computed between groups will be the same whatever the scale. That is:

$$(\bar{Y}_T - \bar{Y}_C) = (\bar{G}_T - \bar{G}_C) = (\bar{g}_T - \bar{g}_C) = (\bar{Y}'_T - \bar{Y}'_C) \qquad (21)$$

The choice of scales will influence the estimate s_y, of course, even where it does not affect the mean difference. Where there are pre-treatment mean differences, then it is inappropriate to use $(\bar{Y}_T - \bar{Y}_C)$; but the question is which of the others to use. Some studies to be included in the meta-analysis may report results with gain scores, others may report residual or covariance adjusted scores. There seems to be no *a priori* reasons for preferring one to the other. It is a choice that the reviewer undertaking a particular meta-analysis must take and should report. Consistency is important. Results on one scale can be converted to the other using:

$$(\bar{G}_T - \bar{G}_C) = (\bar{g}_T - \bar{g}_C) - (1 - b_{y \cdot x})(\bar{X}_T - \bar{X}_C). \qquad (22)$$

Mean differences used for the computation of effect sizes will then all be either $(\bar{Y}_T - \bar{Y}_C)$ or the same variety of adjusted group differences used as an approximation of the final status differences for "initially comparable" groups. The form of the mean difference should be recorded so that any systematic differences in effect size related to the form of its calculation can be revealed.

Choice of Standard Deviation for Scaling Mean Differences

The choice of the standard deviation with which to scale the differences between group means is crucial. Variations in choice can be reflected in substantial differences in effect size. Recording the choice made in each case can allow the investigation of any systematic interaction between the choice and the effect size computed but, unless the relationship is simple, other important relationships with effect size may be obscured.

For most problems, it seems preferable to standardize group mean differences by the standard deviation of the final status variable, not by the standard deviation of some type of gain, change or residual score. The choice of a standardizing metric is hardly trivial. Consider an experimental study in which pretests and posttests were administered and in which no pretest mean differences existed. Suppose further that the pretest-posttest correlation is .75, the posttest mean difference is 10 points and the posttest standard deviation is 15. The effect size, $\underline{\Delta}_y$, in terms of the final status measure is:

$$\Delta_y = \frac{10}{15} = .67.$$

As will be seen below, the standard deviation of residual scores in this instance is $15\sqrt{1-.75^2} = 9.92$. Hence, the effect size in terms of the metric of residual scores is:

$$\Delta_r = \frac{10}{9.92} = 1.01.$$

Obviously the choice of metric makes quite a difference in the calculated effect. Neither calculation is wrong; they merely reflect alternative expressions of the general phenomenon of the experimental results. No rigid rules about which metric is best would be advisable, but the metric of the final status measure seems preferable. Final status (i.e., "posttest score") is a phenomenon more readily perceived and experienced than change or gain; hence, the expression of results on the scale of final status is phenomenologically more meaningful. In addition, there are several ways to measure change or gain that are equally good, or bad (Cronbach and Furby, 1970). "Simple gain," "residual gain," "estimated true gain," and others; each has a different variance and would give a different value of ES. It seems better to avoid them all and standardize group mean differences in terms of final status.

Control Group Standard Deviation on Final Status

Direct Use of Control Group Standard Deviation. Where the standard deviation for a control group on final status scores is available it should be used. The relative effect of treatment with respect to no treatment can then be readily described in terms of the distribution of scores for untreated subjects. Of course, separate effect sizes could be estimated using both control group and experimental group standard deviations. These effect sizes need different interpretations since they express the mean differences in terms of different distributions. The most straightforward procedure is to use the control group distribution as the point of reference. For cases in which the treatment and control group standard deviations are not homogeneous, the treatment group standard deviation will vary with the nature of the treatment. Attempting to keep track of such variations through analysis and interpretation will unnecessarily complicate the analysis.

Retrieval From Standard Deviations on an Adjusted Metric

In the preceding discussion of choice of standard deviation, all standard deviations were taken to be expressed on the metric of the final status scores. If those scores have been adjusted in some way, the standard deviation on the final status metric needs to be retrieved from that on the adjusted metric. Procedures for making such adjustments are described in this section.

Raw Gain Scores. With raw gain score defined by (16) the variance of the raw gain scores can be shown to be:

$$\sigma_G^2 = \sigma_Y^2 + \sigma_X^2 - 2\rho_{xy}\sigma_x\sigma_y \qquad (23)$$

which, if it can be assumed that $\sigma_x = \sigma_y$, reduces to:

$$\sigma_G^2 = \sigma_y^2\{2(1 - \rho_{xy})\} \ . \qquad (24)$$

If the control group standard deviation is provided in terms of raw gain scores as s_G, its standard deviation on the final scores can be obtained from:

$$s_y = \frac{s_G}{\sqrt{2(1 - r_{xy})}} \qquad (25)$$

In many studies reporting in terms of raw gain scores, no information is provided about the correlation between the two status measures. It is also important to note that the correlation required is r_{xy} for the control group or, at least, a pooled within groups estimate of it. If the correlation is not provided, a reasonable guess can probably be made if something is known about the tests involved. For standardized tests, a published test-retest reliability might be appropriate.

Residual Status Scores. With residual status scores defined by (18) the variance of the residual scores can be shown to be:

$$\sigma_g^2 = \sigma_y^2(1 - \rho_{xy}^2) \tag{26}$$

without any necessary assumption about equality of σ_x and σ_y.

If the control group standard deviation is provided in terms of residual scores as s_g, its standard deviation on the final status scores can be obtained from:

$$s_y = \frac{s_g}{\sqrt{1 - r_{xy}^2}} \tag{27}$$

Information about the correlation between scores on the two status measures is more likely to be provided in studies using residual scores than in studies using raw gain scores. The correlation required is the pooled within-group correlation not the control group correlation. Since the residuals are calculated using a pooled estimate of slope, and not separate group estimates, it is with the pooled estimate of correlation that the unreduced standard deviation can be recovered. If the control group standard deviation on residual scores, s_g, is available it should be used rather that a pooled estimate.

Covariance Adjusted Final Status Scores. One effect of covariance adjustments is to reduce the within-group standard deviation in a manner similar to that described for residual scores. If the standard deviation for the control group on the residual scores is given, the standard deviation for the final status scores can be estimated using formula (27).

If only the covariance adjusted pooled within-group mean square, MS'_w, is known a pooled estimate of the within-group standard deviation on final status scores can be obtained from:

$$\hat{s}_Y = \sqrt{\frac{MS'_w}{(1 - r_{xy}^2)} \cdot \frac{(df_w - 1)}{(df_w - 2)}} \tag{28}$$

Retrieval From Higher Order Factorial Designs

Many experimental comparisons of a treatment and a control condition use more complex designs than the simple comparison of two groups. Some introduce other factors into a higher-order analysis of variance design to examine interactions. In the process these designs, create a new definition of within-cell variance. Others introduce stratification of subjects (matching of pairs being an extreme example) to reduce the error variance and obtain a more powerful significance test. The use of repeated measures designs in which subjects are matched with themselves is intended to achieve even more power by the same means.

In reports of studies of this type, only the pooled information in analysis of variance tables is provided. Means must be found to retrieve an appropriate estimate of the control group standard deviation.

Additional Factors of Theoretical Interest. If a higher order analysis of variance is used to explore interactions between the treatment and other factors, that information should not be lost but should instead be coded into the meta-analysis. It is just such interactions that meta-analysis may reveal between studies. Any results which reveal such interactions within studies should be preserved in the data for the meta-analysis. For example, a study to compare treatment and control conditions (Factor A) may stratify the sample of subjects into males and females (Factor B) to study the interaction of the treatment with the subject's gender. For an effect size based on the difference between the overall treatment and control means $(\bar{Y}_{T..} - \bar{Y}_{C..})$ the appropriate standard deviation would be that for the total control group. A pooled estimate of this would be given by:

$$\hat{s}_y = \sqrt{\frac{(SS_B + SS_{AB} + SS_w)}{(df_B + df_{AB} + df_w)}} \quad . \qquad (29)$$

An effect size for males alone would be based on the mean difference $(\bar{X}_{TM.} - \bar{X}_{CM.})$. The appropriate standard deviation would be the one for the control group males for which a pooled estimate would be given by:

$$\hat{s}_y = \sqrt{MS_w} \quad . \tag{30}$$

Stratification on a Continuous Variable Correlated with Outcome. In some studies subjects are stratified on a continuous variable which is correlated with the final status measure. This design allows the within cells sum of squares from the corresponding unstratified design to be partitioned as:

$$SS_{w(A)} = SS_B + SS_{AB} + SS_{w(AB)} \tag{31}$$

as for the case where B is a factor of theoretical interest. Although this design also allows a more powerful test of the treatment effect, there is usually no substantive interest in the between levels variation or the treatment by levels interaction. The control group standard deviation should be obtained as the pooled estimate in formula (29).

If the stratification is achieved by matching pairs, there will be no $SS_{w(AB)}$ term. Only the terms SS_B and SS_{AB} will exist to be pooled. Where the matched pairs data are analyzed by a dependent groups t-test, the standard error of the mean difference between pairs is:

$$\sigma_{\bar{d}} = \sqrt{\frac{\sigma_T^2 + \sigma_C^2 - 2\rho_{TC}\sigma_T\sigma_C}{n}} \tag{32}$$

Where σ_T and σ_C are the standard deviations be the treatment and control groups, ρ_{TC} is the correlation between pairs and n is the number of pairs. If the standard deviations for experimental and control conditions are assumed to be homogeneous, then (32) becomes:

$$\sigma_{\bar{d}} = \sqrt{\frac{2\sigma_y^2}{n}(1 - \rho_{TC})} \tag{33}$$

If the standard error of the mean difference between pairs is reported, the control group standard deviation on the final status measure can be estimated as:

$$s_y = s_{\bar{d}} \sqrt{\frac{n}{2(1 - r_{TC})}} \qquad (34)$$

Since the correlation between pairs, r_{TC}, will probably not be reported it must be estimated. The matching will have been done on some variable X measured before the experiment. The partial correlation of scores on the outcome measure Y between members of pairs, controlling for the common X score for members of each pair, will be:

$$\rho_{Y_T Y_C \cdot X} = \frac{\rho_{Y_T Y_C} - \rho_{Y_T X}\rho_{Y_C X}}{\sqrt{(1 - \rho^2_{Y_T X})(1 - \rho^2_{Y_C X})}} \qquad (35)$$

If the correlation between X and Y is the same for each group, that is ρ_{XY}, then:

$$\rho_{TC \cdot X} = \frac{\rho_{TC} - \rho^2_{XY}}{(1 - \rho^2_{XY})} \qquad (36)$$

and, therefore,

$$\rho_{TC} = \rho^2_{XY} + (1 - \rho^2_{XY})\rho_{TC \cdot X} \quad . \qquad (37)$$

If all that members of a pair have in common can be accounted for by their common scores on the matching variable, then the partial correlation between their scores on any other variable, partialing out their scores on the matching variable, should be zero. A reasonable estimate of the correlation between pairs on the final status measure then would be:

$$\hat{r}_{TC} = r^2_{XY} \qquad (38)$$

If r_{XY} (within group) is not provided in the report, a resonable guess can be made if something is known abut the tests involved.

Stratification on a Continuous Variable of Theoretical Interest. In some studies stratification on a continuous variable may be used to introduce a factor in which there is theoretical interest. For example, in research on ability grouping some studies test only overall mean performances of students taught in homogeneous groups and students taught in heterogeneous groups. Other studies examine, as well, the possibility of differential effectiveness, presenting and testing the significance of differences between homogeneous and heterogeneously grouped students at various levels of ability. Effect sizes can be estimated for both the overall mean differences and the mean differences at different ability levels. The question is, however, which standard deviation should be used to scale the mean differences at specific ability levels--the total control group standard deviation (or a pooled estimate of it), or the standard deviation for the sub-test of the control group at that level (or a pooled estimate of it).

The choice will depend on both the interpretation to be made of the effect sizes and the extent of aggregation of effect sizes. If mean effect sizes over all levels are to be computed, or if effect sizes for various ability levels are to be compared, they should be scaled in terms of the standard deviation of the whole control group. If, from the analysis, it emerges that there are different effect sizes for different ability levels, new effect size estimates based on the control group for each particular level can be calculated. These effect sizes will be indices of the efficacy of treatment at a particular ability level wtih reference to the distribution of the scores of the relevant untreated groups at that level.

If a study presents data for only a part of the total distribution it will be necessary to estimate the standard deviation for the whole control population from the available standard deviation for a truncated section of it. Otherwise the effect sizes calculated will vary according to the homogeneity of the truncated portion used. For this estimation, information will be required about the correlation between the stratifying variable and the final status measure and the selectivity of the sub-group on the grouping variable.

An alternative to estimating the total control group standard deviation, however, would be to use the reported standard deviations and to rate the extent of the truncation of the distribution on a crude three to five point scale. These ratings could be correlated with the effect sizes to determine whether there is any relationship.

Repeated Measures Analyses. Where the treatment and control conditions are such that they can both be applied to the same sample, repeated measures designs are sometimes used to avoid inter-subject variability between groups. In the simplest case, where treatment is one factor (A) and subjects the other (S), the error term for testing the significance of the difference between the treatment group means is the A x S interaction mean square. An estimate of the appropriate control group standard deviation can be obtained if the sums of squares for S and A x S are pooled. Similar approaches to pooling can be used for mixed model designs in which subjects are nested under some additional factors but crossed with treatments.

TRANSLATION OF SIGNIFICANCE LEVELS INTO EFFECT-SIZE

Imagine that in the report of a study it is recorded only that a particular test statistic (e.g., t or F or Fisher's Z-transformation of r) was calculated on n *cases* and that its level of significance (i.e., tail area under the null hypothesis) was p. How can one transform this meager information into a measure of effect size or correlation? Provided that the p-value was reported exactly and not rounded to coarse approximations such as $.05 > p > .01$ (in which case some very crude conventions must be adopted), the transformation is straightforward. If, for example, it is reported that a two group t-test with $n_1 = n_2 = 6$ was significant at the $p = .02$ level (two-tailed test), then it is a simple matter of looking up the value of t in a t-table:

$${}_{.99}t_{10} = 2.76.$$

Thus, one knows n_1, n_2 and the value of the t-test; hence, one can proceed to Δ via the conventional steps derived and illustrated elsewhere:

$$\Delta = t\sqrt{\frac{1}{n_2} + \frac{1}{n_2}}$$

$$= 2.76 \sqrt{\frac{1}{6} + \frac{1}{6}}$$

$$= 1.59 .$$

The reasoning and methods are similar for all of the other test-statistics for which we have derived transformations to r or Δ (see Glass, 1977; Smith, Glass & Miller, 1979; and the first and second quarterly reports). A slight complication may arise at this point. Some investigators attempting an integrative analysis have routinely transformed any $\underline{p}$ value into its corresponding unit normal deviate $\underline{z}$, then into an Δ or $\underline{r}$. The transformation via $\underline{z}$ introduces small errors into the resulting estimates; when the particular test statistic on which $\underline{p}$ is based is known, then it is more accurate to transform via that statistic. For example, in the illustration above with $\underline{p} = .02$ and $\underline{n}_1 = \underline{n}_2 = 6$, the transformation via $\underline{z}$ (which essentially ignores the "degrees of freedom" problem) gives the following estimate of Δ.

$$_{.99}z = 2.326$$

$$\begin{aligned} ES &= z\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 2.326\sqrt{\frac{1}{6} + \frac{1}{6}} \\ &= 1.34 . \end{aligned}$$

The earlier estimate equaled 1.59; the error introduced by transforming via $\underline{z}$ instead of $\underline{t}$ is over 15% of the value of Δ.

Aside from this minor complication, the transformation of $\underline{p}$ values, given $\underline{n}$, into Δ or $\underline{r}$ is rather obvious, and it proceeds by means of conventional statistical tables of significance levels and formulas previously developed for transforming test statistics.

TRANSFORMING NON-PARAMETRIC STATISTICS

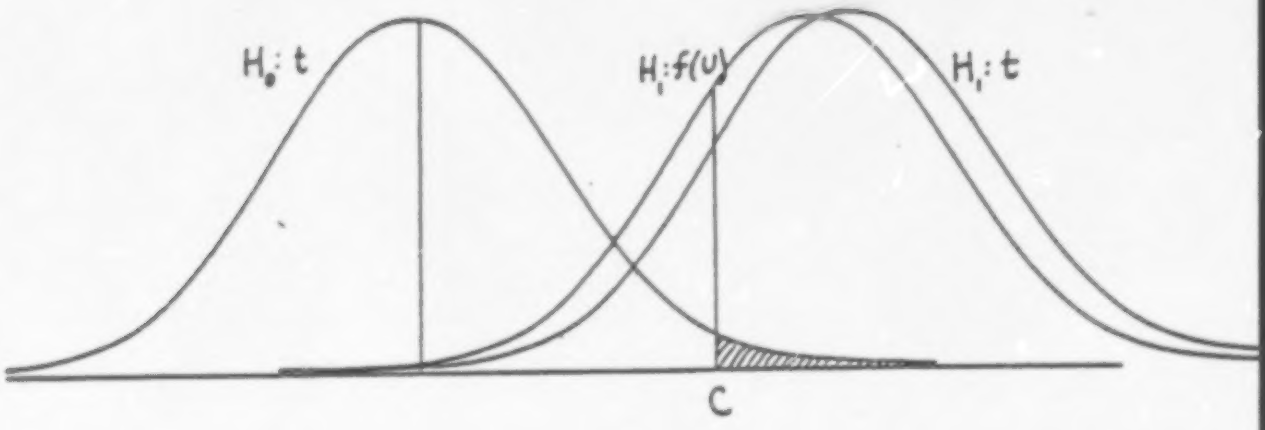
Suppose that a study involved the test of a null hypothesis about equivalent locations of two distributions, and a Mann-Whitney U-test was performed and reported. The U-test competes with a normal-distribution t-test of means in these circumstances; the U-test was once popular because it was believed to be safer when parametric assumptions were violated. The safety proved largely illusory, and today the t-test is the method of choice. But many studies reported U-test results, and it is necessary to consider how information about Δ , say, can be retrieved from them.

No simple transformation of U into Δ is possible since the U-test and most other non-parametric tests do not test simple hypotheses about population means. However, one could substitute for the reported U-statistic the value of t that has the equivalent level of significance. For example, with $n_1 = n_2 = 10$, a $U = 23$ has a two-tailed significance level of $p = .05$. The corresponding t is $t_{.975-18} = 2.10$. From this t-statistic an Δ is found in the conventional manner:

$$\Delta = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= .939 .$$

The above series of transformations appear sensible and adequate, but one refinement may be possible. Nonparametric tests are known to have less power than parametric counterparts where the latter exist. Thus, a U-statistic significant at the $p = .05$ level probably corresponds to a t-statistic that is significant at the .03 or .02 level. For example, it is known that in many circumstances the power of the U-test is about 95% as large as the power of the t-test, a situation illustrated below:



The area to the right of C under the curve H_1:t is p_t, the power of the t-test against the particular alternative hypothesis illustrated. The area above C under H_1:f(U) is p_u, the power of the U-test. It is generally true that $p_u/p_t = 3/\pi$ as $n \to \infty$ (Mood, 1954). Now suppose that p_u is approximately .94 in a particular situation. Then the corresponding power of t is $p_u(\pi/3) = .94(1.0472) = .984$. For large n_1 and n_2, the values of U (appropriately standardized) and t that cut off 94% and 98.4% of the area under roughly normal curves are 1.55 and 2.14. Hence, the small 5% difference in power gives rise to quite large differences in test statistics and, hence, in approximations of Δ's or r's. The prevalence and importance of these differences depend on the relative powers of various non-parametric and parametric tests.

TRANSFORMING DICHOTOMOUS OUTCOME VARIABLES INTO EFFECT SIZES

Experimental outcomes are frequently measured in crude dichotomies where refined metric scales do not exist: dropped out vs. persisted in school, remained sober vs. resumed drinking, convicted vs. not convicted of a crime. It seems inappropriate with such data to calculate means and standard deviations and take a conventional ratio. One approach to this problem is to attempt to recover underlying but unobservable metric (e.g., motivation to stay in school), the experimental and control groups are distributed normally as in Figure 5.4. It is assumed that there exists a cut-off point, C_X, such that if motivation to stay in school falls below C_X, the pupil will drop out. What can be observed are the proportions P_E and P_C of the groups which fall below C_X. Under the normal distributions assumption,

$$p_E = \int_{-\infty}^{z_E} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad (39)$$

where

$$z = \frac{X - \bar{X}_E}{s_E}.$$

Clearly, Z_E is simply the standard normal deviate which divides the curve at the $100P_E$th percentile and can be obtained from any table of the normal curve. Likewise, Z_C is that value of the standard normal variable which cuts off the bottom $100P_C$ percent of the distrubution. Since,

$$z_E = \frac{C_x - \bar{X}_E}{s_E}$$

and

$$z_C = \frac{C_x - \bar{X}_C}{s_C},$$

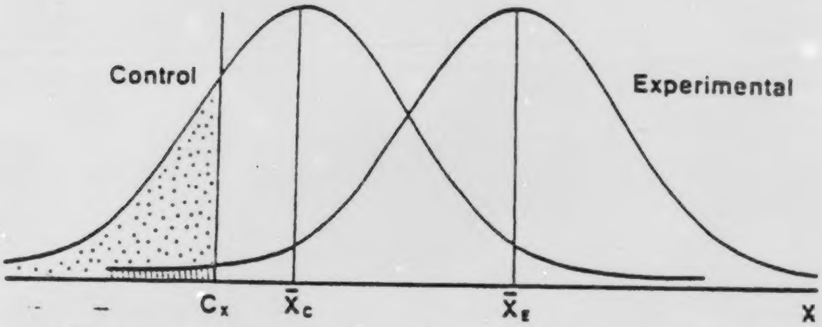


Figure 5.4 Model of the recovery of metric effect-size measures from dichotomous findings.

it can be shown under the assumption of homogeneous variances that

$$z_C - z_E = \frac{\bar{X}_E - \bar{X}_C}{s_x} = \Delta \ .$$

Thus, effect-size measures on hypothetical metric variables can be obtained simply by differencing the standard normal deviates corresponding to the percentages observed in the experimental and control groups. The reasoning followed here is essentially the same as that which underlines probit analysis in biometrics (see Finney, 1971). Where the unobservable metric distributions ought to be assumed skewed in an expected direction, the methods of logit transformation will be more appropriate (Ashton, 1972).

Table 5.6

Probit Transformation of Difference
In Proportions to Effect Size

P_e

P_c	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
.05	0	.36	.60	.80	.97	1.12	1.25	1.39	1.51	1.64	1.77	1.89	2.03	2.16	2.31	2.48	2.68	2.92	3.28
.10		0	.24	.44	.61	.76	.89	1.03	1.15	1.28	1.41	1.53	1.67	1.80	1.95	2.12	2.32	2.56	2.92
.15			0	.20	.37	.52	.65	.79	.91	1.04	1.17	1.29	1.43	1.56	1.71	1.88	2.08	2.32	2.68
.20				0	.17	.32	.45	.59	.71	.84	.97	1.09	1.23	1.36	1.51	1.68	1.88	2.12	2.48
.25					0	.15	.28	.42	.54	.67	.80	.92	1.06	1.19	1.34	1.51	1.71	1.95	2.31
.30						0	.13	.27	.39	.52	.65	.77	.91	1.04	1.19	1.36	1.56	1.80	2.16
.35							0	.14	.26	.39	.52	.64	.78	.91	1.06	1.23	1.43	1.67	2.03
.40								0	.12	.25	.38	.50	.64	.77	.92	1.09	1.29	1.53	1.89
.45									0	.13	.26	.38	.52	.66	.80	.97	1.17	1.41	1.77
.50										0	.13	.25	.39	.52	.67	.84	1.04	1.28	1.64
.55											0	.12	.26	.39	.54	.71	.91	1.15	1.51
.60												0	.14	.27	.42	.59	.79	1.03	1.39
.65													0	.13	.28	.45	.66	.89	1.25
.70														0	.15	.32	.52	.76	1.12
.75															0	.17	.37	.61	.97
.80																0	.20	.44	.80
.85																	0	.24	.60
.90																		0	.36
.95																			0

The transformation of dichotomous information to metric information via probits or logits makes it possible to expand greatly the data base of a meta-analysis. Frequently, studies on a single topic will encompass both metric and dichotomous measurement of outcomes. Having to integrate findings separately by type of outcome measurement is inconvenient as well as less than the broadest, most comprehensive integration of research possible.

Table 5.6 provides the the rapid calculation of Δ given p_e and p_c. For example, suppose that $p_e = .60$ and $p_c = .40$; from the table, the value of Δ is found to be .50. Suppose, as a second illustration that $p_e = .35$ and $p_c = .70$. Then the sign of the effect size will be reversed after referencing Table 5.6 with .70 for columns and .35 for rows: -.91.

Several minor technical problems have arisen in connection with this technique: 1) what should be done when the distributions underlying the dichotomies are not normal?, 2) what if the two distributions (that giving rise to $\underline{p}_e$ and that yielding $\underline{p}_c$) have different variances?, 3) how does the probit transformation compare to treating the dichotomy as an ordered metric and simply calculating $\Delta = (p_e - p_c)/\sqrt{p_c(1 - p_c)}$?, 4) how can a probit transformation be carried out when $\underline{p}$ equals either zero or one?

Non-normality.

We have examined alternative underlying distributions that could serve as a basis of a transformation method like probits. Two distributions seem particularly useful: a) the logistic distribution, and b) the beta distribution. Their probability density distributions are as follows:

Logistic: $P(x) = \{\text{sech}^2[(x-a)/2k]\}/4k$

Beta: $P(x) = [x^{v-1}(1-x)^{w-1}]/B(v,w)$, where $B(v,w)$ is the beta function.

The logistic curve has slightly "thicker tails" than the normal distribution to recommend it, it is a symmetric curve, slightly more peaked in the center and thinner in the intermediate regions than the normal. The following comparison or ordinates makes these features clear:

	z-score				
Ordinate of	-4	-3	-2	-1	0 . . .
Normal	.0001	.0044	.0540	.2420	.3989 . . .
Logistic	.0013	.0078	.0458	.2186	.4535 . . .

Although these differences in ordinates appear small, they yield large differences in estimated effects when transformed first to percentiles then to z-scores.

The beta distribution is a large family of curves bounded between 0 and 1 for the variate x and encompassing symmetric and asymmetric curves of widely varied shapes. The beta distribution for $v = 4$ and $w = 2$ is depicted below.

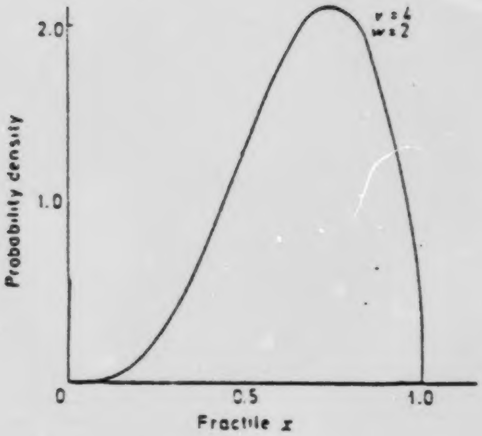


Figure 5.5 Probability density function for the beta variate β: v, w.

By changing v and w, the beta distribution can be given any desire skewness. Thus, it is a useful distribution for describing asymmetric variables. Furthermore, its percentiles have been extensively tabluated (Pearson and Hartley, 1962).

We applied, where appropriate, probit transformations and metric calculation of effect sizes on a body of literature in drug therapy and psychotherapy. The discrepancy between the average effect sizes for the two different methods proved to be relatively large, as Table 5.7 below reveals.

It must be emphasized that the comparison in Table 5.7 is based on two sets of data not necessarily equivalent in all important respects. However, the direction of the difference (favoring the probit transformation by nearly two-tenths standard deviation units) is consistent with the expectation that violations of the normality assumption of the probit method are likely to inflate effect-size estimates, particularly where dichotomies are extreme (.95 vs. .05 or worse).

Table 5.7

Comparison of Average Effects Calculated by Either Probit Transformation or Metric Statistics From 112 Experiments on Drug and Psychotherapy

Method	No. of Δ	Average Effect Size, Δ
Probit Transformation	53	.651
Metric Statistics	351	.494

Heterogeneous Variances. Suppose that one observes p_e as the proportion of cases exceeding some fixed point, C, on a scale of measurement for which Z_e is normally distributed with mean and standard deviation μ_e and σ_e. The quantity p_c is similarly defined with Z_c having mean and standard deviation μ_c and σ_c. Now if p_e and p_c are transformed into the unit normal deviates, z_e and z_c, that cut off the upper $100p_e\%$ and $100p_c\%$ of the normal curve, then:

$$z_e = \frac{C - \mu_E}{\sigma_E} \quad \text{and} \quad z_c = \frac{C - \mu_c}{\sigma_c} \quad .$$

It is easily shown that:

$$z_c - z_e(\sigma_e/\sigma_c) = \Delta = \frac{\mu_E - \mu_C}{\sigma_c} \quad ,$$

the mean difference standardized against the control group standard deviation. If one knew the value of σ_e/σ_c or had a good hunch about it, then Δ could be easily calculated by weighted z_e by the ratio σ_e/σ_c. But it is more realistic (because σ_e/σ_c will nearly always be unknown) and important to ascertain how Δ is affected if σ_e and σ_c are unknown and heterogeneous. Beginning with $z_c - z_e$ and permitting σ_e and σ_c to differ, one quickly arrives at the expression:

$$z_c - z_e = \frac{C(\sigma_e - \sigma_c)}{\sigma_e\sigma_c} + \frac{\sigma_c\mu_e - \sigma_e\mu_c}{\sigma_c\sigma_e} \qquad (40)$$

It is interesting to note that this expression depends on C, the hypothetical cut-off point used in determining "success" in both the experimental and control groups. The equation has not worked out to any form that is particularly neat or useful. There is probably little point in pursuing it much further. It is sufficient merely to record that heterogeneous variances affect the probit transformation both through their effect on the mean difference and the value of the criterion score. One is advised to be alert to the possibility of unequal variances and to use a transformation such as $z_c - z_e(\sigma_e/\sigma_c)$ when possible.

Probits vs. Dichotomous Variables. It has occurred to some to ask whether the probit transformation of two dichotomies is roughly equivalent to treating the dichotomies as merely a limiting case of an effect size from the manifest variable, e.g.,

$$\Delta'_d = \frac{p_e - p_c}{\sqrt{p_c(1 - p_c)}} \qquad (41)$$

This expression is simply the mean difference between the two dichotomies standardized by the standard deviation of the control group.

The appropriate question to ask is how closely this formulation agrees with the effect size calculated from the probit transformation, viz.,

$$\Delta_p = z_c - z_e \text{ , where}$$

z_e is the unit normal deviate that marks off the upper $(100p_e)\%$ of the area under the normal curve, and z_c is similarly defined. The ratio of Δ_p to Δ_d for various values of p_e and p_c is easily calculated. Values of the ratio for p_e ranging from .1 to .9 in steps of .10 are tabulated below:

Values of the Ratio Δ_p/Δ_d

		p_e, Proportion of Successes in the Experimental Group								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
	.1	--	1.32	1.14	1.04	0.96	0.92	0.90	0.91	0.96
p_c	.2	1.76	--	1.27	1.20	1.12	1.10	1.09	1.12	1.21
	.3	1.74	1.46	--	1.29	1.20	1.19	1.20	1.25	1.38
Proportion	.4	1.70	1.47	1.38	--	1.19	1.24	1.25	1.33	1.49
f successes	.5	1.60	1.40	1.31	1.22	--	1.27	1.31	1.40	1.60
in the	.6	1.50	1.34	1.27	1.19	1.24	--	1.33	1.44	1.68
ntrol group	.7	1.33	1.25	1.20	1.17	1.20	1.24	--	1.46	1.74
	.8	1.21	1.12	1.09	1.09	1.12	1.18	1.27	--	1.76
	.9	0.96	0.91	0.90	0.92	0.96	1.03	1.14	1.32	--

These ratios are disconcertingly large, in most cases. For example, if $p_e = .20$ and $p_c = .10$, the effect size calculated from the probit transformation is nearly one-third larger than the effect calculated from treating the data as a manifest dichotomy. It seems clear that in spite of the problems of non-normality and heterogeneous variances that may plague the probit transformation, the calculation of effects from dichotomies without consideration of underlying distributions is not an acceptable alternative.

Probits at the Extremes. A vexing problem with probit transformations from dichotomous to metric data arises when n cases reveal either 0 or n "successes." Then the proportion $p = f/n$ equals either 0 or 1, and the corresponding unit normal deviates are infinite ($-\infty$ and $+\infty$). Consider a typical example. Ten experimental subjects are treated for dyslexia, and at the end of six months each reads sufficiently well to be promoted ($p_e = 10/10 = 1$). None of the ten control groups is promoted ($p_c = 0/10 = 0$). The corresponding unit normal deviates are $z_e = +\infty$ and $z_c = -\infty$, and $\Delta = \infty-(-\infty) = 2\infty$! Absurd. Suppose that it was decided arbitrarily to change one case in each sample to avoid this problem. Then p_e would be taken equal to 9/10 and p_c to 1/10. Now the unit normal deviates are 1.282 and -1.282, respectively; and $\Delta = 2.564$. Suppose a compromise between 0 and 1 "success" was struck at 0.5 so that p_c equaled $0.5/10 = .05$ and, similarly, $p_e = .95$. The resulting value of Δ is $1.645-(-1.645) = 3.290$. The difference between 3.290 and 2.564 is too large to ignore; and the difference of either from ∞ is too gruesome to contemplate. A method is needed for dealing non-arbitrarily with p's of 1 or 0. One solution is afforded by Bayesian statistics.

We shall assume that p is a sample estimate of π where $p = \frac{x}{n}$ and x is binomially distributed. The Bayesian posterior distribution of π is given by:

$$Pr(\pi|x) = \frac{Pr(\pi)\ Pr(X|\pi)}{Pr(X)} ,$$

where $Pr(\pi)$ is the prior distribution of π assumed to be uniform on the interval 0 to 1.

Now Pr(x) is given by:

$$\Pr(x) = \int_0^1 \Pr(\pi) \binom{n}{x} \pi^{x} (1 - \pi)^{n-x} \, d\pi \qquad (42)$$

Since Pr (Π) is a constant k, and recognizing that the terms in π integrate to a Beta distribution, formula (42) becomes

$$\Pr(x) = k \binom{n}{x} B(x + 1, n - x + 1),$$

where $B(u, v) = \left[\Gamma(u) \quad \Gamma(v)\right] / \Gamma(u + v),$ where

$\Gamma(u) = (u - 1)! = (u - 1)(u - 2) \ldots 3 \cdot 2 \cdot 1,$ when u is an integer.

The distribution of X given Π is simply the binomial:

$$\Pr(X \mid \Pi) = \binom{n}{x} \Pi^{x} (1 - \Pi)^{n-x}.$$

Thus, the posterior distrubution of Π is given by:

$$\Pr(\Pi \mid X) = \frac{k \binom{n}{x} \Pi^{x} (1 - \Pi)^{n-x}}{k \binom{n}{x} B(x + 1, n - x + 1)}.$$

The Bayesian estimate of Π , denoted by $\hat{\Pi}$, is the mean of the posterior distribution:

$$E(\Pi \mid x) = \hat{\Pi} = \int_0^1 \frac{\Pi \; . \; \Pi^{x} (1 - \Pi)^{n-x} \, d\hat{\Pi}}{B(x + 1, n - x + 1)}$$

$$= \frac{B(x + 2, n - x + 1)}{B(x + 1, n - X + 1)}$$

$$= \frac{\Gamma(x + 2) \; \Gamma(n - x + 1) \; \Gamma(n + 2)}{\Gamma(n + 3) \; \Gamma(x + 1) \; \Gamma(n - x + 1)} = \frac{x + 1}{n + 2}$$

This result is the important one: *assuming a uniform prior distribution for π, the Bayesian estimate of π, the binomial parameter, equals $\hat{\pi} = (X + 1)/n + 2$ where n is the sample size and X is the observed number of successes.* (Solutions are also possible for various non-uniform prior distributions of π, especially the Beta distribution, for example.)

This result offers a non-arbitrary method of resolving difficulties of probit transformation for the cases of $p = 1$ or 0. If $X = 0$ in a binomial sample of n, then whereas $p = 0$, the Bayesian estimate $\hat{\pi}$ equals $(0 + 1)/(n + 2)$. Likewise, at the other end of the scale of p of 1 corresponds to a $\hat{\pi}$ of $(n + 1)/(n + 2)$. For example, in the illustration discussed earlier, $p_e = 10/10$ would yield $\hat{\pi}_e = 11/12 = .92$; and $p_c = 0/10$ would give $\hat{\pi}_c = 1/12 = .08$. Hence Δ equals $1.40-(-1.40) = 2.80$. This solution seems non-arbitrary and reasonable. Having found it, we see no reason why it should not be applied across the board, that is, regardless of the value of $p = X/n$, if a uniform prior distribution of π is reasonable, the, $\hat{\pi}$ should be taken to equal $\hat{\pi} = (X + 1)/(n + 2)$.

An interesting problem arises when one's purposes are study integration. Suppose that ten separate studies of five persons each yielded identical results, one of five "successes." Each value of p would equal 1/5, and the average of all the p's or the pooled value across the ten studies would both equal .20. However, the average of the Bayesian estimates would be $(\hat{\pi}_1 + \ldots + \hat{\pi}_5)/5 = 5 \cdot (2/7)/5 = .29$. The Bayesian correction in small samples can be substantial, even though in a pooled sample it would be insignificant, e.g., $\hat{\pi}_{\text{pooled}} = 11/52 = .21$ vs. $10/50 = .20$. Thus the average of many small sample Bayesian estimates can be quite different from a pooled Bayesian estimate. A pooled estimate would seem preferable, but pooling obviates the examination of study-to-study variation in findings, which is much in the spirit of our approach to integrating research.

OUTCOMES OF CORRELATIONAL STUDIES

In the meta-analysis of correlational studies, one is integrating correlation coefficients descriptive of the relationship between two variables, such as achievement and socioeconomic level, or teacher personality and pupil learning. The quantitative description of findings from correlational studies presents fewer complications than do experimental studies.

Illustrations of the integrative analysis of correlational studies will be drawn from a study of the relationship between pupils' socioeconomic status (SES) and their academic achievement. White (1976) collected over 600 correlation coefficients from published and unpublished literature. The coefficients were analyzed to determine how their magnitude was related to varying definitions of SES, different types of achievement, age of the subjects, and so on. White found that the 636 available correlations of SES and achievement averaged .25 with a standard deviation of about .20 and positive skew. Thus, SES and achievement correlation is below what is generally believed to be the strength of association of the two variables. The correlation diminished as students got older, $\underline{r}$ decreasing from about .25 at the primary grades to around .15 late in high school. SES correlated higher with verbal than math achievement (.24 vs. .19 for 174 and 128 coefficients, respectively). When White classified the SES and achievement correlations by the type of SES measure employed (see Table 5.8), SES measured as income correlated more highly with achievement than either SES measured by the education of the parents or the occupational level of the head of household.
Several reliable trends in the collection of 600 coefficients could help

methodologists designing studies and sociologists constructing models of the schooling-social system.

It probably matters little whether analysis is carried out in the metric of r_{xy}, r^2_{xy} or Fisher's Z transformation of r_{xy}. The final results ought to be expressed in terms of the familiar r_{xy} scale, however. There appears to be no good reason to transform r_{xy} to Fisher's Z at the intermediate stages of aggregation and analysis, though this is sometimes recommended. Fisher's transformation was developed to solve an inferential problem, and it would be an unlikely happenstance if it proved to be the method of choice for combining correlation measures from several studies. It is frequently recommended that two or more r_{xy}'s be squared, averaged, and the square root taken rather than averaged directly. However, it is fairly easy to show that the choice seldom makes a practical difference. A little algebra applied to the ratio of $(r_1 + r_2)/2$ to $\sqrt{(r_1^2 + r_2^2)/2}$ will show that the discrepancy between the two depends primarily on the size of the difference between r_1 and r_2 and that they must be enormously different for the two averaging methods to differ in any important way. For example, the three coefficients -- .20, .30, and .40 -- average .30 directly; and they average .31 if first squared and averaged, and the square root is taken. A gap of approximately more than .50 between r_1 and r_2 is needed to separate $(r_1 + r_2)/2$ and $\sqrt{(r_1^2 + r_2^2)/2}$ by more than .05. The researcher can safely decide whether the scale of r_{xy} or r^2_{xy} is more meaningful to him and work in that metric throughout an integration of correlational studies.

The correlational studies referred to here deal with ordinal, metric variables. Correlational results which involve genuine dichotomies or polychotomies (e.g., sex, ethnic group) should be recast into more

Table 5.8

Average Correlation between SES and Achievement for Different Kinds of SES Measure

SES Measure	Average r_{xy}*
Indicators of parents' income	.315 (19)
Indicators of parents' education	.185 (116)
Indicators of parents' occupation level	201 (65)

* Number of coefficients averaged in parentheses.

informative descriptive measures such as standardized differences among means, and the techniques of "effect-size" measurement discussed above may then be applied. Where the two variables correlated are conceived of as having metric properties -- even if the technology of measurement at the time fell short of actual metric measurement -- then one ought to seek to transform all correlation measures to the scale of Pearson's product-moment correlation coefficient.

When a large field of correlational research is collected, a bewildering variety of statistics is encountered: biserial and point-biserial correlation coefficients, rank-order correlations, phi coefficients, contingency coefficients, contingency tables with chi square tests, _t_-tests, analyses of variance, and more. In White's analysis of SES and achievement correlation, a variety of methods of reporting what was basically a correlational finding was encountered. Of 143 studies, 37 reported _t_ or _F_ statistics, 71 reported Pearson _r_'s, 8 reported chi square or non-parametric statistics, and 27 presented only graphs or tables of means.

There usually is an algebraic path from the reported statistics to a Pearson correlation coefficient or an approximation to one. Some signposts along the paths are set out in Table 5.9, where it is indicated how one might travel from particular forms of reported data to a product-moment correlation measure.

Table 5.9

Guidelines for Converting Various Summary Statistics Into Product-Moment Correlations

Reported Statistic	Transformation to r_{xy}	References
a) Point-biserial correlation, r_{pb}	$^{*}r_{xy} = r_{pb}\sqrt{n_1 n_2}/(un)$ u = ordinate of unit normal distribution n = total sample size	Glass and Stanley (1970, p. 171)
b) $t = \frac{X_1 - X_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$	$r_{pb} = \sqrt{\frac{t^2}{t^2 + (n_1 + n_2 - 2)}}$ then convert r_{pb} to r_{xy} via a) above.	Glass and Stanley (1970, p. 318)
c) t based on extreme groups.	$\rho \cong \frac{t(\sqrt{2/n})}{\sqrt{\frac{{}^{4}z^{2}}{p^2} + (\frac{z^2}{p^2} - \frac{{}^{x}z}{p})t^2(\frac{2}{n})}}$ n = within cell n. p = proportion cut at each end. z = ordinate on normal curve at the cut. x = standard normal denote corresponding to p (abscissa value)	Based on Feldt, Psychometrica, 1971, p. 315. Rearranged by Glass.
d) $F = MS_b/MS_w$ for $J = 2$ groups.	$\sqrt{F} = \|t\|$ then proceed via b) above.	
e) $F = MS_b/MS_w$ for $J > 2$ groups.	1) Collapse J groups to 2 & then proceed via d) above, or 2) $r_{xy} = \eta = \sqrt{SS_b/(SS_b + SS_w)}$	Hays (1973, pp. 683-684)
f) χ^2 only (i.e., no frequencies reported) for a contingency table.	$^{**}r_{xy} \cong P = \left(\frac{\chi^2}{\chi^2 + n}\right)^{1/2}$ n = total sample size.	Kendall & Stuart (1967, p. 557 ff)

Table 5.9 Continued

Reported Statistic	Transformation to r_{xy}	References
g) 2 x 2 contingency table.	Calculate tetrachoric r_{xy} from tables	Glass and Stanley (1970, p. 165 ff)
h) R x C contingency table.	Collapse to a 2 x 2 table and proceed via g) above.	
i) Spearman's rank correlation, r_s.	$r_{xy} = r_s$ since the translation of r_s to r_{xy} under bivariate normality is nearly a straight line.	Kruskal (1958)
j) Mann-Whitney U.	Transform U to r-rank-biserial via $r_{rb} = 1 - 2U/(n_1 n_2)$.	Willson (1976)

* $r_{xy} \cong 1.25 r_{pb}$ when $p = n_1/n$ is between .2 and .8 (Magnusson, 1966, p. 205).

** P is Pearson's coefficient of contingency and $P^2 \to p^2$ as the number of categories in the table increases. With few categories, the estimate can be unduly low.

Another common instance of transforming results involves converting a correlation, r, into a standardized mean difference. For example, Coleman's survey of equality of educational opportunity reported a correlation coefficeint between class-size, X, and achievement, Y. But most other studies reported the relationship in terms of means and variance on achievment for particular class-sizes, leading to the measure Δ_{S-L} described in first section of this report. Knowing only r_{xy} and $\overline{X}$ and s_x, the measure Δ_{S-L} can be calculated assuming a normal distribution of X and a linear relationship of X and Y. Values of S and L must be specified on X; they can be arbitrarily designated as any two convenient percentiles, e.g., P and 100-P. Then $S = \overline{X} - zs_x$ and $L = \overline{X} + zs_x$, where z is the unit normal deviate at the percentile 100-P.

From r_{xy}, we can calculate the regression line of Y on X from

$$b_{yx} = r_{xy}(s_y/s_x), \text{ and}$$

$$b_0 = \overline{Y} - b_{yx}\overline{X}.$$

The mean values of $\overline{Y}$ corresponding to S and L are calculated by substitution into the regression equation. The within group variance on $\overline{Y}$ is simply the variance error of estimate, known to equal $s_y^2(1 - r^2)$. Combining these facts leads to

$$\Delta_{S-L} = 2zr_{xy}(1 - r_{xy}^2)^{-\frac{1}{2}}, \text{ where}$$

z is the unit normal deviate at the Pth percentile of the normal cureve (S being at the Pth percentile in the distribution of X and L being at the 100-Pth percentile of X).

The above conversion seems unobjectionable, and surely is provided that X is roughly normally distributed and the regression of Y and X is linear. However, when Y has a curvilinear regression on X, the value of Δ_{S-L} will be somewhat in error.

NONPARAMETRIC MEASURE OF EXPERIMENTAL EFFECT

Kraemer and Andrews (1980) have recently devised a descriptive measure of effect size that appears to have advantages over traditional standardized mean difference measures. Their measure is based on frequency statistics and the inverse normal transformation. The most important property of the Kraemer-Andrews measure is that it is invariant with respect to monotonic transformations of the dependent variable. As this is written, it is too soon to evaluate the utility of this new measure, but early reactions seem promising.

CHAPTER SIX

METHODS OF ANALYSIS

The analysis of data in a meta-analysis is properly approached as an instance of multi-variate data analysis in which the studies are the units on which measurements are taken and the study characteristics (Chapter Four) and findings (Chapter Five) are the many variables. The point of having come this far in our treatment of meta-analysis is the belief that the import of many studies described in many ways cannot be grasped by the reader without the aid of techniques of arranging, ordering, relating -- in short, without the help of statistical methods. Univariate description; frequency tabulations, correlations, linear model estimation, regression analysis, factor analysis, analysis of covariance, discriminant function analysis -- any of the methods of statistical analysis that have proved to be useful in extracting meaning from data are potentially useful in meta-analysis. One's attitude toward the data may be exploratory (Tukey, 1977) or confirmatory, descriptive or inferential; it doesn't matter. We are breaking no new ground here. We are merely illustrating the application of well-known statistical methods in a context in which researchers are prone to forget that they are as useful, indeed necessary, as in other familiar contexts.

In this chapter, we shall first deal briefly with the simple univariate descriptive analysis of study findings. Then we shall describe methods of examining the correlation of study findings and characteristics. Third, the estimation of treatment effects where study findings can be arranged in the manner of factorial experiments will be investigated. Fourth, attention will be given to the special possibilities of integrating study findings where both the independent and dependent variables are measured on quantitative scales. Fifth, problems of statistical inference as they apply in meta-analysis will be discussed.

SIMPLE DESCRIPTION OF STUDY FINDINGS

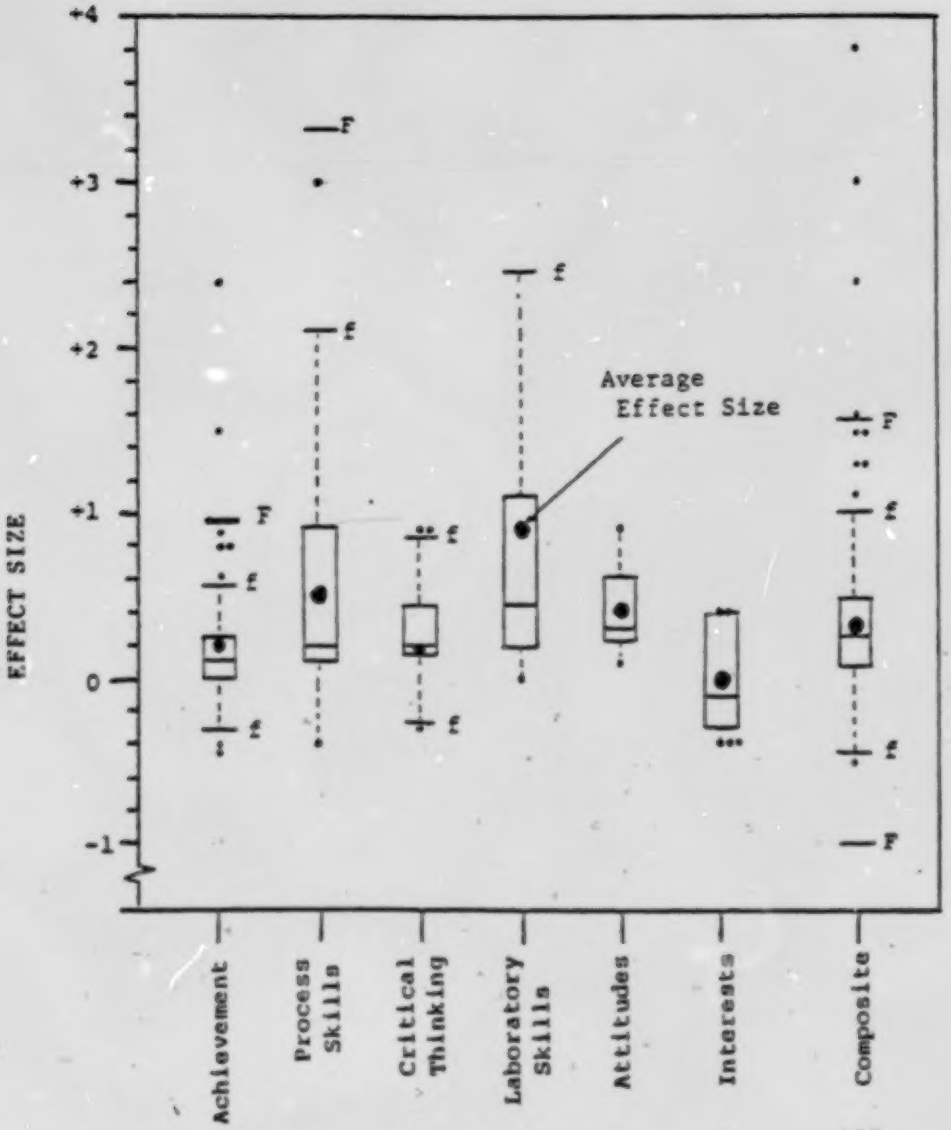
Once the findings of the studies in a meta-analysis have been measured (whether by means of an effect size, a correlation coefficient or whatever), all the standard methods of tabulating and describing statistics may be usefully applied: frequency distributions, averages, measures of variability, and the like. In this respect, we much prefer Tukey's (1977) innovative and ingeneous methods of exploratory data analysis to the unimaginative lot of techniques presented in most statistical methods textbooks. An illustration might help the reader understand our preference.

El-Nemr (1979) found 59 experimental studies in which were compared traditional teaching of biology and biology taught as a process of inquiry. These studies yielded nearly 250 effect size measures in which inquiry-teaching was compared with traditional teaching of biology. The effect size measures seven categories descriptive of type of outcome: science achievement, science process skills, critical thinking skills, laboratory skills, attitudes toward the biology course, interest in science, and "composite" (an average of the preceding outcomes). Plots of the characteristics of the distributions of effect sizes for each outcome category appear as Figure 6.1.

Consider the first category of outcomes in Figure 6.1. The 59 experiments yielded 30 effect sizes based on the measurement of achievement (since achievement was not measured in every experiment). Each effect size is of the form

$$\triangle_{I - T} = \frac{\overline{Y}_I - \overline{Y}_T}{s_T}$$

The distribution of the 39 achievement effect sizes is described by the lines, letters and dots above "Achievement" in Figure 6.1. The basic descriptive technique is the "box-and-whisker" plot with auxilliary features. The central box or rectangle marks off the "hinges" (roughly, the first and third quartiles) of the distribution of effect sizes and the median (ordinary definition) as the sizes lie between the top and the bottom of the box with 25 percent of those inside the box on either side of the median. The hinges for the achievement effect sizes are at .02 and .23, approximately, and the median is at .17. The large black dot inside the box indicates the location of the average of the 39 effect sizes; for achievement, the mean is above the median. The dotted line emanating from both ends of the box measures the distance to the "inner fence," a distance arbitrarily chosen to be one-and-one-half times the length of the box (i.e., 150% of the hinge range). The lower-case letter f marks the inner fence. Data points that lie outside the inner fence are "outliers," and each is denoted by a small dot. At the same distance beyond the inner fence that the inner fence lies beyond the ends of the box one marks off the "outer fence" with an upper-case F. Data points beyond the outer fence are "far outliers." One casts a suspicious eye at outliers and looks with even greater incredulity on far outliers. They may represent oddities (measurement reporting errors, misprints,



n:	39	30	19	7	9	9	133
Medians:	.17	.29	.23	.55	.29	-.13	.23
Means:	.20	.50	.18	.87	.38	-.01	.32

Figure 6.1. Summary statistics for effect sizes in seven classes of outcome from comparisons of inquiry vs. traditional teaching of biology. (After El-Nemr, 1979)

miscalculations, and whatever) that ought to be eliminated or given different weight in describing the typical features of the data.

Notice, for example, that among the 39 achievement effect sizes in Figure 6.1 there are four outliers and two far outliers. If the two far outliers are eliminated and the average effect size recalculated, the average drops from .20 to .10. The median drops a little, but less than the 50 percent drop for the mean. Consider the "Process Skills" outcome category. Here, a substantial discrepancy exists between the median and the mean with the latter one and two-thirds times larger than the former. But the mean is probably distorted by the single far outlier of 3.0; removing this outlier drops the mean to .41, because of the positive skew in the data for process skills shown by the fact that the median is far closer to the lower hinge than the upper. Generally the means are larger than the medians, except for "Critical Thinking" where the order is reversed. And although the inquiry approach to teaching biology was superior to traditional teaching in most respects, it was no better at stimulating pupils' interest in science.

Correlating Study Characteristics and Findings

The next step beyond the simple description of study findings is the study of the relationship between study characteristics and findings. This second stage of analysis is addressed to such questions as whether the findings are homogeneous for all types of subject (e.g., person) or whether they are positive for some types of subject and negative for others, whether the findings are strong when viewed with certain research methods (e.g., subjective outcome appraisals), whether the short-term findings differ substantially from the long-term results and so forth.

Any one of the many statistical techniques for studying the association or relationship between two variables may find useful application at this stage: contingency table analysis, regression analysis, correlation analysis with its many subspecies (e.g., Pearson's r, point-biserial or biserial correlation, curvilinear correlation). Since study findings will be measured on a metric scale (Δ , r, etc.), metric measures of relationship deriving from Pearson product-moment notions will be the most powerful and useful.

Consider an illustration. In their first meta-analysis of the effects of psychotherapy, Smith and Glass (1977) compiled several hundred effect size measures for nearly four hundred controlled outcome evaluations. Among the characteristics of the studies coded were the following:

	Characteristics	Coding
1)	Organization of therapy	1 = individual, 2 = group.
2)	Duration of therapy	No. of hours.
3)	Years experience of therapist	No. of years.
4)	Client diagnosis	1 = psychotic, 2 = neurotic
5)	IQ of clients	1 = low, 2 = medium, 3 = high
6)	Age of clients	Age in years.
7)	Social-economic-cultural similarity of therapist & clients	1 = very similar, . . ., 4 = very dissimilar.
8)	Internal validity of study	1 = high, 2 = medium 3 = low.
9)	Date of publication of study	Year
10)	"Reactivity" of outcome measure	1 = low, 2 = low ave., 3 = ave., 4 = high ave., 5 = high.
11)	No. of months after therapy of outcome measurement	No. of months

Each of the eleven study characteristics was correlated with the effect size. The linear correlation coefficients obtained are reported in Table 6.1.

Table 6.1

Correlations of Several Descriptive Variables with Effect Size

Variable	Correlation with effect size
Organization (1 = individual; 2 = group)	−.07
Duration of therapy (in hours)	−.02
Years' experience of therapists	−.01
Diagnosis of clients (1 = psychotic; 2 = neurotic)	.02
IQ of clients (1 = low; 2 = medium; 3 = high)	.15**
Age of clients	.02
Similarity of therapists and clients (1 = very similar; . . . ; 4 = very dissimilar)	−.19**
Internal validity of study (1 = high; 2 = medium; 3 = low)	−.09*
Date of publication	.09*
"Reactivity" of outcome measure (1 = low; . . . ; 5 = high)	.30**
No. of months posttherapy for follow-up	−.10*

* $p < .05$.
** $p < .01$.

The correlations are generally low, although several are reliably non-zero. Some of the more interesting correlations show a positive relationship between an estimate of the intelligence of the group of clients and the effect of therapy, and a somewhat larger correlation indicating that therapists who resemble their clients in ethnic group, age, and social level get better results. The effect sizes diminish across time after therapy as shown by the last correlation in Table 6.1, a correlation of - .10 which is closer to -.20 when the curvilinearity of the relationship is taken into account. The largest correlation is with the "reactivity" or subjectivity of the outcome measure. The multiple correlation of the eleven study characteristics with the effect size was equal to about .50; thus, 25 percent of the variance in study findings can be accounted for by variations in the characteristics of the studies. There is not space here to pause and consider the many implications of the relationships reported in Table 6.1; in this example, they are numerous, and they have not escaped either those who comment on the benefits of psychotherapy or those who concern themselves with the methodology of its evaluation (see Chapter Seven for further discussion of this point).

A more controversial use of the relationships of study characteristics to findings involves the attempt to equate various classes of studies and then observe comparative results. Imagine a simple hypothetical example. Either medication or hypnotherapy can be prescribed for asthmatic children. A set of 50 controlled experiments on the effects of medication show an average effect size of .75; 60 experiments with hypnotherapy give an average effect size of .40. It is observed, however, that on the average the medication experiments measured effects one month after treatment whereas the hypnotherapy experiments measured outomes at six months. Furthermore, within each class of experiment, the regression coefficient of Δ onto "follow-up time" is about the same:

medication: $\hat{\Delta} = .83 - .08$ (No. of months)

Hypnotherapy: $\hat{\Delta} = .65 - .08$ (No. of months)

If the effects of both treatments are estimated for follow-up times of one month, the .35 difference in the uncorrected average comparison (.34 = .75 - .40) shrinks to .75 - .57 = .18 standard deviation units difference between the means of the treatment and control groups. If the regression of effect onto follow-up time were heterogeneous in the regressions slopes between the two therapies, the estimated order of superiority could change from one follow-up time to another.

In our analysis of psychotherapy effects, the regression of effect size onto ten independent variables was performed separately within three quite different classes of psychotherapy: psychodynamic, systematic desensitization, and behavior modification. The results of the three multiple regression analyses appear in Table 6.2.

Table 6.2

Regression Analyses Within Therapies

Independent variable	Unstandardized regression coefficients Psychodynamic (n = 94)	Systematic desensitization (n = 212)	Behavior modification (n = 129)
Diagnosis (1 = psychotic; 2 = neurotic)	.174	−.193	.041
Intelligence (1 = low; . . . ; 3 = high)	−.114	.201	.201
Transformed age[a]	.002	−.002	.002
Experience of Therapist × Neurotic	−.011	−.034	−.018
Experience of Therapist × Psychotic	−.015	.004	−.033
Clients self-presented	−.111	.287	−.015
Clients solicited	.182	.088	−.163
Organization (1 = individual; 2 = group)	.108	−.086	−.276
Transformed months posttherapy[b]	−.031	−.047	.007
Transformed reactivity of measure[c]	.003	.025	.021
Additive constant	.757	.489	.453
Multiple R	.423	.512	.509
σ_e	.173	.386	.340

[a] Transformed age = (Age − 25)(|Age − 25|)$^{\frac{1}{2}}$.
[b] Transformed months posttherapy = (No. months)$^{\frac{1}{2}}$.
[c] Transformed reactivity of measure = (Reactivity)$^{1.[illegible]}$.

Relatively complex forms of the independent variables were used to account for interactions and nonlinear relationships. For example, years' experience of the therapist bore a slight curvilinear relationship with outcome, probably because more experienced therapists worked with more seriously ill clients. This situation was accommodated by entering, as an independent variable, "therapist experience" in interaction with "diagnosis of the client." Age of client and follow-up date were slightly curvilinearly related to outcome in ways most directly handled by changing exponents. These regression equations allow estimation of the effect size a study shows when undertaken with a certain type of client, with a therapist of a certain level of experience, etc. By setting the independent variables at a particular set of values, one can estimate what a study of that type would reveal under each of the three types of therapy. Thus, a statistically controlled comparison of the effects of psychodynamic systematic desensitization, and behavior modification therapies can be obtained in this case. The three regression equations are clearly not homogeneous; hence, one therapy might be superior under one set of circumstances and a different therapy superior under others. A full description of the nature of this interaction is elusive, though one can illustrate it at various particularly interesting points.

In Figure 6.2 estimates are made of the effect sizes that would be shown for studies in which simple phobias of high-intelligence subjects, 20 years of age, are treated by a therapist with 2 years experience and evaluated immediately after therapy with highly subjective outcome measures.

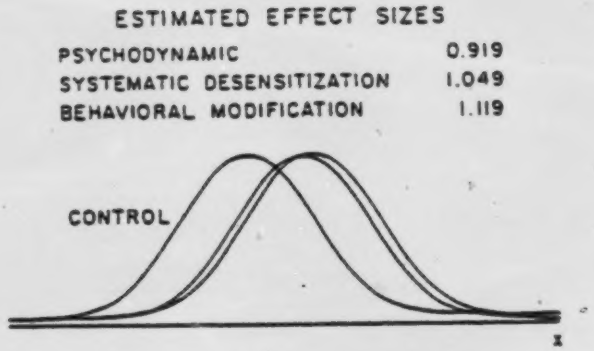


Figure 6.2. Three within-therapy regression equations set to describe a prototypic therapy client (phobic) and therapy situation.

This verbal description of circumstances can be translated into quantitative values for the independent variables in Table 6.2 and substituted into each of the three regression equations. In this instance, the two behavioral therapies show effects superior to the psychodynamic therapy.

In Figure 6.3 a second prototypical psychotherapy client and situation are captured in the independent variable values, and the effects of the three types of therapy are estimated. For the typical 30-year-old neurotic of average IQ seen in circumstances like those that prevail in mental health clinics (individual therapy by a therapist with 5 years experience), behavior modification is estimated to be superior to psychodynamic therapy, which is in turn superior to systematic desentization at the 6-month follow-up point.

Besides illuminating the relationships in the data, the quantitative techniques described here can give direction to future research. By fitting regression equations to the relationship between effect size and the independent variables descriptive of the studies and then by

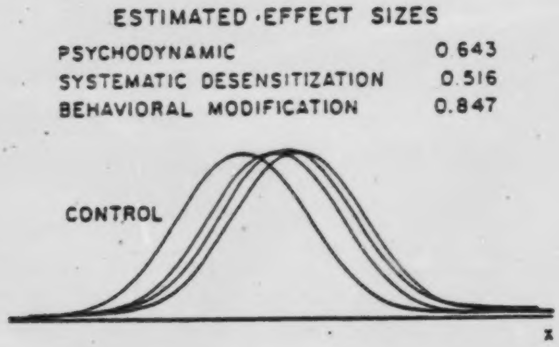


Figure 6.3. Three within-therapy regression equations set to describe a prototypic therapy client (neurotic) and therapy situation.

placing confidence regions around these hyperplanes, the regions where the input-output relationships are most poorly determined can be identified. By concentrating new studies in these regions, one can avoid the accummulation of redundant studies of convenience that overelaborate small areas.

Linear ANOVA Models for Estimation of Effects

Collections of experiments often present odd arrays of comparison to one who wishes an integrated summary of effects. For example, an integration of reading instruction research would encounter experiments comparing Initial Teaching Alphabet (ITA) and Traditional Orthography (TO), other experiments comparing ITA and Diacritical Marking (DM), and still a third type of experiment in which TO and DM are compared. For each comparison, a standardized mean contrast can be calculated (e.g., $\Delta = (\overline{X}_{ITA} - \overline{X}_{TO})/s_X$); but the integration of these various Δ's into a estimation of the effects of the three individual instructional methods is not immediately obvious. One fruitful approach is *via* "effects coding" and the general linear model. For example, the following model can be postulated:

$$\Delta = \beta_{ITA}X_1 + \beta_{TO}X_2 + \beta_{DM}X_3 + e \qquad (1)$$

The variables X_1, X_2 and X_3 take on the values, 1, 0, and -1. If, for example, a particular Δ is based on an experimental comparison of ITA and TO, then $X_1 = 1$, $X_2 = -1$ and $X_3 = 0$. In this way, many Δ's can be regressed onto the X's; and the β's, which are individual effects of the instructional methods, can be estimated.

The technique of "control referencing" that was dealt with briefly in Chapter Five can be approached more conveniently through use of the linear effects models of this section. Suppose, for example, that there exist n experiments in which treatment A is compared to a control group, n experiments in which B is compared with a control group and n experiments in which A and B are compared directly without a control group. There are, thus, three types of effect size measure: Δ_A, Δ_B and Δ_{A-B}. A simple modification of the general linear model like that in (1) above suffices to describe the effects:

$$\Delta = \beta_A X_1 + \beta_B X_2 + e. \qquad (2)$$

$X_1 = 1$ if Δ is of the form A vs. Control,

$X_2 = 1$ if Δ is of the form B vx. Control

$X_1 = +1$ and $X_2 = -1$ if Δ is of the form A vs. B.

For the equal n's example, the data, the design and the parameter matrices are as follows:

$$\begin{bmatrix} \Delta_A \\ \cdot \\ \cdot \\ \cdot \\ \Delta_A \\ \Delta_B \\ \cdot \\ \cdot \\ \cdot \\ \Delta_B \\ \Delta_{A-B} \\ \cdot \\ \cdot \\ \cdot \\ \Delta_{A-B} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \\ 1 & -1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \\ \beta_B \end{bmatrix} + e$$

Denoting the design matrix by $\underline{X}$, the least-squares estimates of the effect parameters are given by

$$\hat{\beta} = (X^T X)^{-1} X^T \Delta$$

The form of $(X^T X)^{-1}$ and $X^T \Delta$ are as follows:

$$(X^T X)^{-1} = \frac{1}{n} \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}, \quad X^T \Delta = \begin{bmatrix} \Sigma\Delta_A + \Sigma\Delta_{AB} \\ \Sigma\Delta_B - \Sigma\Delta_{A-B} \end{bmatrix}$$

Therefore, the estimates of the aggregate effect sizes for treatments A and B are given by

$$(X^TX)^{-1}X^T\Delta = \begin{bmatrix}\hat{\beta}_A \\ \hat{\beta}_B\end{bmatrix} = \begin{bmatrix} 1/3(2\bar{\Delta}_A + \bar{\Delta}_B + \bar{\Delta}_{A-B}) \\ 1/3(2\bar{\Delta}_B + \bar{\Delta}_A - \bar{\Delta}_{A-B}) \end{bmatrix} ,$$

Where the bar above the delta indicates simple average.

A related, but slightly more complex, problem involves treatment components which can be evaluated separately or in combination in experiments. Consider, for example, the treatment of psychological disorders by either drugs or psychotherapy or both.

The experimental literature on drug and psychotherapy addressed the estimation of the separate and interactive effects of drugs and psychotherapy in a variety of ways. The variety is a nuisance. Several types of experiments can be identified which inform one about the drug effect alone, or the drug plus the interaction effect, or the psychotherapy plus the drug plus the interaction effect, and so on in various combinations. An experiment that compares clients' progress under drugs with a group of clients receiving a placebo or nothing estimates the simple drug effect. Whereas an experiment that compares two groups of clients one of which receives drugs-plus-psychotherapy and the other of which receives only drugs provides an estimate of the psychotherapy plus the interaction effect, since one group has the possible advantage of the separate psychotherapy effect and any benefits that result from combining drugs and psychotherapy. Denote the drug effect in isolation when compared with a placebo or no treatment by δ; denote the separate psychotherapy effect by Ψ; and denote the interaction effect of the two by η. Then the comparison of drug therapy and placebo in an experiment estimates δ. The comparison of drug-plus-psychotherapy with psychotherapy estimates $\delta + \eta$

because both sides of the comparison have equal psychotherapy effects. In Table 6.3 appear the possible experimental comparison of drug and psychotherapy and what effects these comparisons estimate.

By arranging and averaging the results from experiments of the six different types specified in Table 6.3, the separate and interactive effects of drug and psychotherapy can be estimated. The organization of data and unknown parameters in Table 6.3 can be viewed as a system of six sources of information and three unknown parameters. Least-squares estimates of the parameters can be calculated by ordinary methods.

Table 6.3

The Structure of Experiments on the Effects of Drug and Psychotherapy

Treatments Compared in the Experiment	Effects Estimated by the Comparison
A. Drug vs. Placebo (or No Treatment)	δ
B. Psychotherapy vs. Placebo	Ψ
C. (Drug & Psychotherapy) vs. Placebo	$\delta + \Psi + \eta$
D. (Drug & Psychotherapy) vs. Drug	$\Psi + \eta$
E. (Drug & Psychotherapy) vs. Psy	$\delta + \eta$
F. Drug vs. Psychotherapy	$\delta - \eta$

If one wished to maintain a distinction between placebo and no-treatment control groups, there would be twelve lines in Table 6.3 instead of six and the structure of effects would change slightly; for example, a Drug vs. No-Treatment experiment would estimate the drug plus the placebo effect since the expectancy effect of administering the drug to the experimental group would not be counter-balanced by an expectancy effect for the no-treatment control group.

In a meta-analysis of psychotherapy research, the question was addressed of the main and interactive effects of psychotherapy and drug therapy. A total of 112 studies was collected, each of which addressed the question in part with one or more experimental comparisons. These 112 studies yielded 566 effect-size measures (i.e., standardized mean differences). For example, a study in which drug treatment was compared with combined drug and psychotherapy treatment, a standardized mean difference of the following form would result: $\Delta = (\bar{X}_{D+P} - \bar{X}_D)/s_X$. In Table 6.4 appear the actual average effect sizes calculated from the findings of the 112 experiments.

As an example of how Table 6.4 can be interpreted, consider the first line of entries. A total of 55 comparisons in the 112 studies involved contrasting the scores of persons who received psychotherapy with those who received no treatment or, at most, a placebo. Such comparisons estimate the magnitude of the psychotherapy effect, ψ ; the estimate equals .30, i.e., the psychotherapy groups averaged three-tenths standard deviation superior to the control groups on the outcome variables. Consider as a second example the 94 comparisons of drug-plus-psychotherapy with psychotherapy alone. Such comparisons estimate the separate drug effect, δ , and the interactive effect, η , which results when drug and psychotherapy are combined in the same treatment. The psychotherapy effect, ψ , is

Table 6.4

Average Effect Sizes from Various Experimental Comparisons

Made in the Experiments on Drug and Psychotherapy

Comparison	Parameter(s) Estimated	Average Δ	No. of Δ's
Psychotherapy vs. No-Treatment or Placebo	ψ	.30	55
Drug Therapy vs. No-Treatment or Placebo	δ	.51	351
Drug & Psychotherapy vs. Drug	$\psi + \eta$	.41	10
Drug & Psychotherapy vs. Psychotherapy	$\delta + \eta$	.44	94
Drug vs. Psychotherapy	$\delta - \psi$	.10	7
Drug & Psychotherapy vs. No-Treatment or Placebo	$\delta + \psi + \eta$	.65	49

Note. ψ denotes the separate or "main" effect of psychotherapy;
δ denotes the separate effect of drug therapy; and
η denotes their interaction.

not reflected in the contrast because it is present on both sides of the comparison. The 94 effect sizes which estimate $\delta + \eta$ have an average of .44 . The remainder of the table can be understood in like manner.

From simple inspection, it appears that the drug effect of .51 is more than half again as large as the psychotherapy effect of .30 . The interaction effect is slightly more difficult to comprehend from merely inspecting the entries in Table 6.4. That the drug-plus-psychotherapy vs. drug comparison, which estimates $\psi + \eta$, if a full one-tenth standard deviation larger than the .30 estimate of ψ from the first line of the table might lead one to believe that η is positive; but the comparison of the estimates of $\delta + \eta$ and δ (being .44 and .51, respectively) reverses this impression. Inspection is too arbitrary and confusing. Several comparisons in the table contain information about the same parameters; it seems reasonable that every source of information about a parameter should be used in estimating it. A complete and standard method of combining the data in Table 6.4 into estimates of the parameters is needed. Such a method is suggested when one recognizes that the two middle columns of Table 6.4 constitute a system of linear equations, three of them independent and containing three unknowns (ψ, δ and η). The method of least-squares statistical estimation can be applied to obtain estimates of the separate and interactive effects of drug and psychotherapy.

The data and parameters of Table 6.4 can be written as a set of simultaneous linear equations as follows:

$$\begin{bmatrix} .30 \\ .51 \\ .41 \\ .44 \\ .10 \\ .65 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \psi \\ \delta \\ \eta \end{bmatrix}$$

Denoting the vector of data by Δ and the design matrix by X, the solution for the parameter estimates is as follows:

$$(X^T X)^{-1} X^T \Delta = \begin{bmatrix} \hat{\psi} \\ \hat{\delta} \\ \hat{\eta} \end{bmatrix} \qquad (3)$$

$$(X^T X)^{-1} = \begin{bmatrix} 1/2 & 1/4 & -1/2 \\ 1/4 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1 \end{bmatrix}, \text{ and}$$

$$X^T \Delta = \begin{bmatrix} 1.26 \\ 1.70 \\ 1.50 \end{bmatrix}$$

Hence, the estimates of the parameters are found from $(X^TX)^{-1}X^T\Delta$ to be

$$\hat{\psi} = .31$$

$$\hat{\delta} = .42$$

$$\hat{\eta} = .02$$

Each effect is expressed on a scale of standard deviation units. Thus, the data of Table 6.4 lead to the conclusion that with the groups of clients studied psychotherapy produces outcomes that are about one-third standard deviation superior to the outcomes from placebo or untreated control groups. The drug effect is only about a third greater than the psychotherapy effect. An effect of $.31s_x$ will move an average client from the middle of the control group distribution to about the 62nd percentile; an effect of .42 would move the average client to only about the 66th percentile.

INTEGRATING STUDIES THAT HAVE QUANTITATIVE INDEPENDENT VARIABLES

Many bodies of research literature involve the examination of the relationship between dependent and independent variables, both described quantitatively. Where the quantitative character of the independent variable can be preserved, the gain in precision of the integration of findings can be considerable. Examples of problems where this is true include class-size and achievement, the duration of effects of any treatment, study time and achievement, and countless laboratory problems in the social sciences. Consider, for example, a research integration problem faced by Underwood (1957) in his work on memory. Over fifteen studies were available to him addressed to the question of the efficiency of recall as a function of the ordinal position of the items to be recalled in a series of lists. Underwood plotted the curve reproduced below as Figure 6.4 and concluded that efficiency of recall was largely a function of interference from items

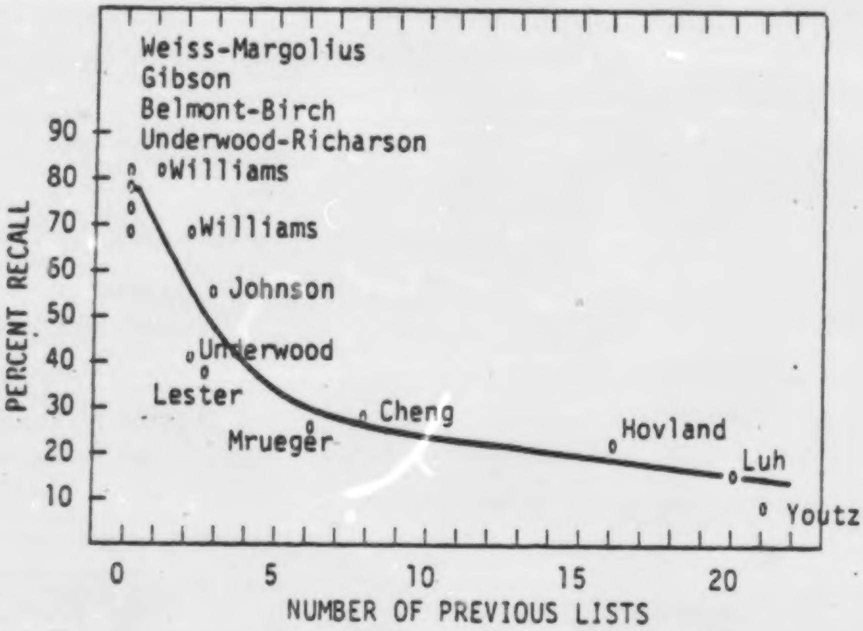


Figure 6.4. Recall as a function of previous lists learned as determined from a number of studies.

previously memorized. The curve in Figure 6.4 represents a simple problem in research integration; it could be fit adequately with a logarithmic curve or many other alternatives to a straight line. But the problems presented by many other quantitative independent and dependent variables are more complex. Consider the relationship between class-size and educational achievement.

A Modification of Multiple Linear Regression

A simple statistic is desired that describes the relationship between class-size and achievement as determined by a study. No matter how many class-sizes are compared, the data can be reduced to some number of paired comparisons, a smaller class against a larger class. Certain differences in the findings must be attended to if the findings are later to be integrated. The most obvious differences involve the actual sizes of "smaller" and "larger" classes and the scale properties of the achievement measure. The actual class-sizes compared must be preserved and become an essential part of the descriptive measure. The measurement scale properties can be handled by standardizing all mean differences in achievement by dividing by the within group standard deviation (a method that is complete and discards no information at all under the assumption of normal distributions). The eventual measure of relationship seems straightforward and unobjectionable:

$$\Delta_{S-L} = \frac{\overline{X}_S - \overline{X}_L}{\hat{\sigma}} ,$$

where:

$\overline{X}_S$ is the estimated mean achievement of the smaller class which contains $\underline{S}$ pupils;

$\overline{X}_L$ is the estimated mean achievement of the larger class which contains $\underline{L}$ pupils; and

$\hat{\sigma}$ is the estimated within-class standard deviation, assumed to be homogeneous across the two classes.

As a first approximation to studying the class-size and achievement relationship, it is considered irrelevant that the particular types of achievement that lie behind the variable $\underline{X}$ are quite different knowledges and skills measured in quite different ways.

If distributional assumptions about $\underline{X}$ are needed to add meaning to particular values of Δ_{S-L}, normality will be assumed. For example, suppose $\Delta_{S-L} = +1$. Then assuming normal distributions within classes, the average pupil in the smaller class scores at the 84th percentile of the larger class. These interpretations are occasionally helpful, but seldom critical, and our investment in the normality assumption is not great. It would be no surprise nor any concern if the assumption proved to be more or less wrong, and it's probably not far off in most instances.

There exist several alternative statistical techniques for integrating a large set of Δ_{S-L}'s so as to describe the aggregated findings on the class-size and achievement relationship. A large, square matrix could be constructed in which the rows and columns are class-sizes and the cell entries are average values of Δ_{S-L}; nearly equal values of average deltas could be connected by lines to form "iso-deltas" in much the manner as economic equilibrium curves are used to depict three-variable relationships. Or a variation of psychometric scaling could be employed: a square matrix of class-sizes could be constructed for which each cell entry would be the proportion of times the row class-size gave achievement greater than the column class-size. This matrix could be scaled by means of Thurstone's Law of Comparative Judgment, which would locate the class-sizes along an achievement continuum. (This method was used and the results were reasonably satisfactory.) Finally, regression equations could be constructed in which Δ_{S-L} is partitioned into a weighted linear combination of $\underline{S}$ and $\underline{L}$ and functions thereof and error. There is much to recommend this latter procedure, and the technique eventually employed is a variation of it. But the regression of Δ_{S-L} onto only $\underline{S}$ and $\underline{L}$ requires three dimensions to be depicted. Anything more complex than a simple two-dimensional curve relating achievement to the size of class was considered undesirably complicated and beyond the easy reach of most audiences who hold a stake in the results.

The desire to depict the aggregate relationship as a single-line curve is confounded with the problem of essential inconsistencies in the design and results of the various studies. A single study of class-size and achievement may yield several values of Δ_{S-L}. In fact, if $\underline{k}$ different class-sizes are compared on a single achievement test, $k(k-1)/2$ values of Δ_{S-L} will result. This set of Δ's from a single study will form a consistent set of values in that they can be joined to form a single connected graph depicting the curve of achievement as a function of class-size. However, various values of Δ_{S-L} arising from different studies can show confusing inconsistencies. For example, suppose that Study #1 gave Δ_{10-15}, Δ_{10-20}, and Δ_{15-20}, and Study #2 gave Δ_{15-30}, Δ_{15-40}, and Δ_{30-40}. A few moments reflection will reveal that there is no obvious or simple way to connect these values into a single connected curve.

The eventual solution to these problems proceeded as follows: Δ_{S-L} was regressed onto a quadratic function of $\underline{S}$ and $\underline{L}$ by means of the least-squares criterion; then that set of values of $\hat{\Delta}$ that could be expressed as a single, connected curve was found.

The regression model selected accounted for variation in Δ_{S-L} by means of S, S^2 and L. Obviously, something more than a simple linear function of S and L was needed, otherwise a unit increase in class-size would have a constant effect regardless of the starting class-size S; and the S^2 term seemed as capable of filling the need as any other. The size differential between the larger and smaller class, L-S, was used in place of L for convenience. Thus, the Δ_{S-L} values were used to fit the following model:

$$\Delta_{S-L} = \beta_0 + \beta_1 S + \beta_2 S^2 + \beta_3 (L-S) + \varepsilon$$

Fitting this model by least-squares will result in the curved regression surface

$$\hat{\Delta}_{S-L} = \hat{\beta}_0 + \hat{\beta}_1 S + \hat{\beta}_2 S^2 + \hat{\beta}_3 (L-S) \tag{4}$$

The problem now is to find the set of $\hat{\Delta}$'s in this surface that can be depicted as a single curved-line relationship in a plane. The property that must hold for a set of $\hat{\Delta}$'s before they can be depicted as a connected graph in a plane is what might be called the *consistency property*:

$$\Delta_{n_1-n_2} + \Delta_{n_2-n_3} = \Delta_{n_1-n_3}$$

for $n_1 < n_2 < n_3$. If this property is not satisfied, then one is in the strange situation of claiming that the differential achievement between class-sizes 10 and 20 is not the sum of the differential achievement from 10 to 15 and then from 15 to 20.

When the consistency property is imposed on (4), it follows that:

$$\begin{aligned}&\hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_2-n_1) + \hat{\beta}_0 + \hat{\beta}_1 n_2 + \hat{\beta}_2 n_2^2 + \hat{\beta}_3 (n_3-n_2)\\ &= \hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_3-n_1)\end{aligned} \tag{5}$$

Simple algebraic reduction of (5) produces the following:

$$\hat{\beta}_0 + \hat{\beta}_1 n_2 + \hat{\beta}_2 n_2^2 = 0 \tag{6}$$

The two solutions to the quadratic equation in (6) are points n_2 such that if $\hat{\Delta}_{S-L}$ is measured with n_2 as either the larger, L, or smaller, S, class-size, then the resulting set of $\hat{\Delta}$'s will lie on the four dimensional regression curve in (4) but can be depicted as a single line curve in a plane. Since n_2 becomes the point around which values of n_1 and n_3 are selected, it will be called the pivot point.

A Logarithmic Model

The above modified regression approach for integrating studies with quantitative independent variables is disappointingly complex. Fortunately we have found two simpler alternatives: 1) a logarithmic model and 2) a non-linear model.

The logarithmic model can be illustrated with the class-size problem.

Assume that the Δ for a comparison of class-size 1 and any other class-size C has the form

$$\Delta_{1-C} = \beta \log C + e, \quad \text{where } e \sim (0, \sigma_e^2).$$

Now consider the values of C denoted by S and L which stand in the relationship $S < L$. Then,

$$\Delta_{1-S} = \beta \log S + e, \text{ and}$$
$$\Delta_{1-L} = \beta \log L + e.$$

Assuming, quite reasonably that

$$\Delta_{S-L} = \Delta_{1-L} - \Delta_{1-S}, \text{ one has that}$$
$$\Delta_{S-L} = \beta \log (S/L) + e. \tag{7}$$

Thus, the parameter β can be estimated by simple least-squares regression of Δ_{S-L} onto $\log(S/L)$. Then a single curve depicting the relationship of Δ to C can be drawn in a plane defined by the two axes C and Z, the (in the calculus sense) of Δ. We have applied this model in the analysis of class-size and achievement with very satisfactory results. It fit the data with lesser mean-square error than did the linear regression approach described above. Furthermore, this simple logarithmic model presents far more tractable problems of statistical inference than the modified regression model.

A Non-Linear Model

A third alternative exists. Its comparative advantages will be pointed out later.

Suppose that a study of the relationship of class-size and achievement is done in which achievement is compared in classes of size n_1, n_2 and n_3. The average achievement in each group is $\overline{Y}_1$, $\overline{Y}_2$ and $\overline{Y}_3$. A simple model for the relationship between achievement and class-size in this study could take the following form:

$$\overline{Y}_i = \sigma\mu\beta^X + \varepsilon .$$

The parameter μ represents a hypothetical level of achievement at class-size zero (i.e., $X = 0$). The parameter σ is an arbitrary scale of measurement parameter. If β is restricted to the interval 0 to 1, then the curve described is an expoential that does not drop off as fast as the logarithmic curve. For example, the following table shows the decay in achievement as class-size increases when $\beta = .90$.

Table 6.5

Comparison of Non-Linear and Logarithmic Models

X	$\mu\beta^{X}$	Based on $\log_e X$
0	μ	
1	$.90\mu$	
2	$.81\mu$	μ
4	$.66\mu$	$.50\mu$
8	$.43\mu$	$.33\mu$
16	$.19\mu$	$.25\mu$
32	$.03\mu$	$.20\mu$

In the third column above, the rate of decay for the logarithmic model is given for comparison. As can be seen, the non-linear model drops off much less rapidly for small values of $\underline{X}$.

The non-linear model can easily be adapted for integrating many different studies by allowing μ and σ to vary depending on the study. By introducing a coding variable w_j which equals 1 when study $\underline{j}$ is considered and zero otherwise, the following integrative model is obtained:

$$\bar{Y}_{ij} = w_j \left[\sigma_j \mu_j \beta^{X_i}\right] + \varepsilon \qquad (8)$$

This integrative non-linear model has 2J + 1 unknown parameters and J · K data points, provided that each study has K means; if at least one study has three means, the model parameters can be estimated by means of non-linear least-squares analysis.

The logarithmic model in (7) would fit data well where the drop off was severe for small values of the quantitative indpendent variable. But the log model has no asymptote, which is often a disadvantage. The non-linear model in (8) would fit data well where the initial drop was less severe, but where an asymptote was approached for large values of X. It ought to be possible to combine the two models additively into a mixed model and gain the benefits of each.

The Logaritmic Model Illustrated

Consider an illustration from research on class-size and achievement. Fourteen experiements were found in which pupils were randomly assigned to classes of different sizes. These fourteen studies yielded over 100 separate comparisons of achievement in smaller and larger classes. The multiplicity of findings is due partly to the fact that in one study there may exist several pairs of class sizes and partly to the fact that a single pair of class sizes may have been measured on more than one achievement test. The latter multiplicity was averaged out and the former retained in the summary of 30 data points in Table 6.6.

One might expect class-size and achievement to be related in something of an exponential or geometric fashion--reasoning that one pupil with one teacher learns some amount, two pupils learn less, three pupils learn still less, and so on. Furthermore, the drop in learning from one to two pupils could be expected to be larger than the drop from two to three, which in turn

Table 6.6

Data on the Relationship of Class-size and Achievement from Studies Using Random Assignment of Pupils.
(Outcomes scaled with $\hat{\sigma} = (s_L + s_S)/2$.)

Study Number	Size of Smaller Class	Size of Larger Class	$\log_e(L/S)$	Δ_{S-L}
1.	25.	1.	In 25.0 = 3.22	.32
2.	3.	1.	In 3.0 = 1.10	.22
2.	25.	1.	In 25.0 = 3.22	1.52
2.	25.	3.	In 8.3 = 2.12	1.22
3.	35.	17.	In 2.1 = .72	-.29
4.	112.	28.	In 4.0 = 1.39	-.03
5.	2.	1.	In 2.0 = .69	.36
5.	5.	1.	In 5.0 = 1.61	.52
5.	23.	1.	In 23.0 = 3.14	.83
5.	5.	2.	In 2.5 = .92	.22
5.	23.	2.	In 11.5 = 2.44	.57
5.	23.	5.	In 4.6 = 1.53	.31
6.	30.	15.	In 2.0 = .69	.17
7.	23.	16.	In 1.4 = .36	.05
7.	30.	16.	In 1.8 = .63	.04
7.	37.	16.	In 2.3 = .84	.08
7.	30.	23.	In 1.3 = .27	.04
7.	37.	23.	In 1.6 = .48	.04
7.	37.	30.	In 1.2 = .21	0
8.	28.	20.	In 1.4 = .33	.15
9.	50.	26.	In 1.9 = .65	.29
10.	32.	1.	In 32.0 = 3.46	.65
11.	37.	15.	In 2.5 = .90	.40
11.	60.	15.	In 4.0 = 1.38	1.25
11.	60.	37.	In 1.62 = .48	.65
12.	8.	1.	In 8.0 = 2.08	.30
13.	45.	15.	In 3.0 = 1.10	.07
14.	14.	1.	In 14.0 = 2.64	.72
14.	30.	1.	In 30.0 = 3.40	.78
14.	30.	14.	In 2.14 = .76	.17
			1.42	.38

$$\Delta_{S-L} = \frac{\bar{X}_S - \bar{X}_L}{(s_S + s_L)/2}$$

n = 14 studies

N = 30 comparisons

is probably larger than the drop from three to four, and so on. A logarithmic curve represents one such relationship:

$$y = \alpha - \beta \log_e C + \epsilon \text{, where} \qquad (9)$$

C denotes class-size.

In formula (9), α represents the achievement for a "class" of one person, since $\log_e 1 = 0$, and β represents the speed of decrease in achievement as a class-size increases. The general curve is graphed in Figure 6.5.

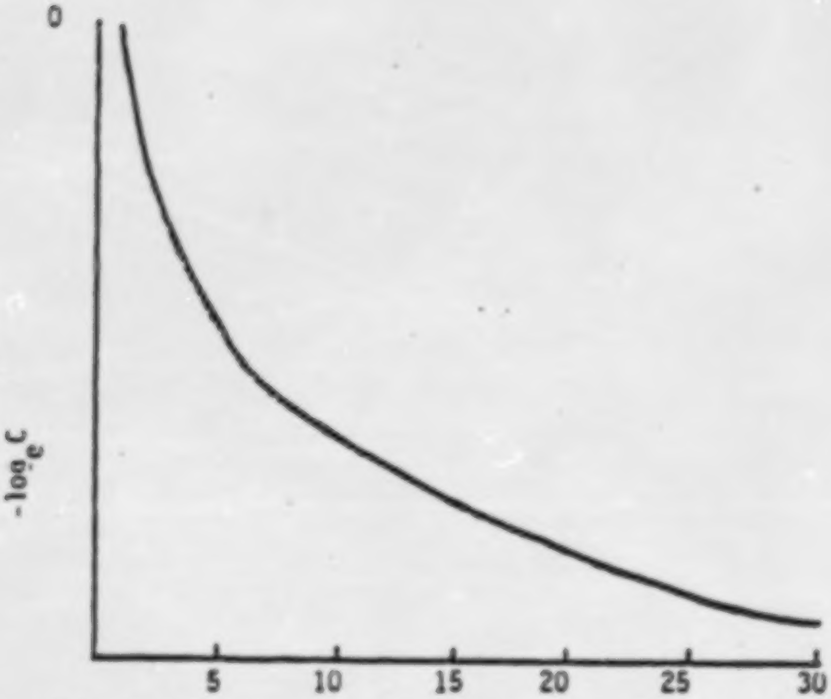


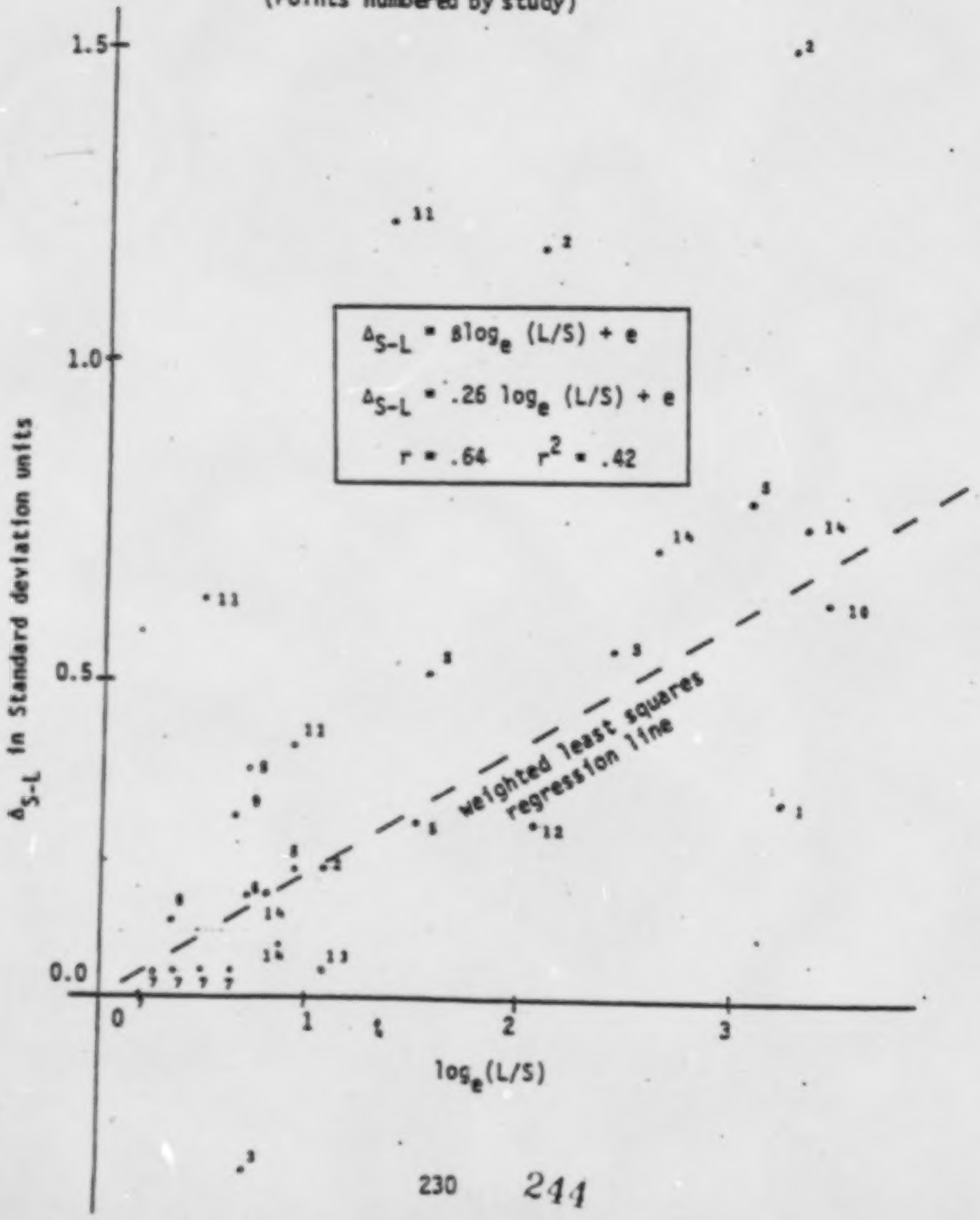
Figure 6.5 Graph of the log curve for the model in formula (9).

Formula (9) can not be fitted to data directly because Y is not measured on a common scale across studies. This problem can be circumvented by calculating Δ_{S-L} for each comparison of a smaller and a larger class

Figure 6.6

Scatter Diagram of Δ_{S-L}

Graphed Against Log_e (L/S).
(Points numbered by study)



within a study. Then, from formulas (7) and (9) one has

$$\begin{aligned}\Delta_{S-L} &= (\alpha - \beta \log_e S + \varepsilon_1) - (\alpha - \beta \log_e L + \varepsilon_2) \\ &= \beta(\log_e L - \log_e S) + \varepsilon_1 - \varepsilon_2 \\ &= \beta \log_e (L/S) + \varepsilon. \qquad (10)\end{aligned}$$

The model in formula (10) is particularly simple and straightforward. The values of Δ_{S-L} are merely regressed onto the logarithm of the ratio of the larger to the smaller class-size, forcing the least-squares regression line through the origin.

$$\hat{\beta} = \frac{\Sigma(\Delta_{S-L})(\log_e L/S)}{\Sigma(\log_e L/S)^2}. \qquad (11)$$

A scatter diagram of the data in Table 6.6 appears as Figure 6.6, in which Δ_{S-L} is graphed against $\log_e(L/S)$. The estimate of β for these data equals 0.27. The value of r is .64, and $r^2 = .42$. The resulting curve relating class-size C to achievement in standard-score units appears as Figure 6.7.

One can either weight each Δ_{S-L} in Table 6.6 equally in deriving an estimate of β, or it can be reasoned that each of the fourteen studies should receive equal weight so that each Δ_{S-L} is multiplied by $2/(k^2-k)$ when it is derived from a study involving k different class-sizes. The estimate of β from the regression involving weighted Δ's is equal to 0.21, which agrees closely with the earlier result.

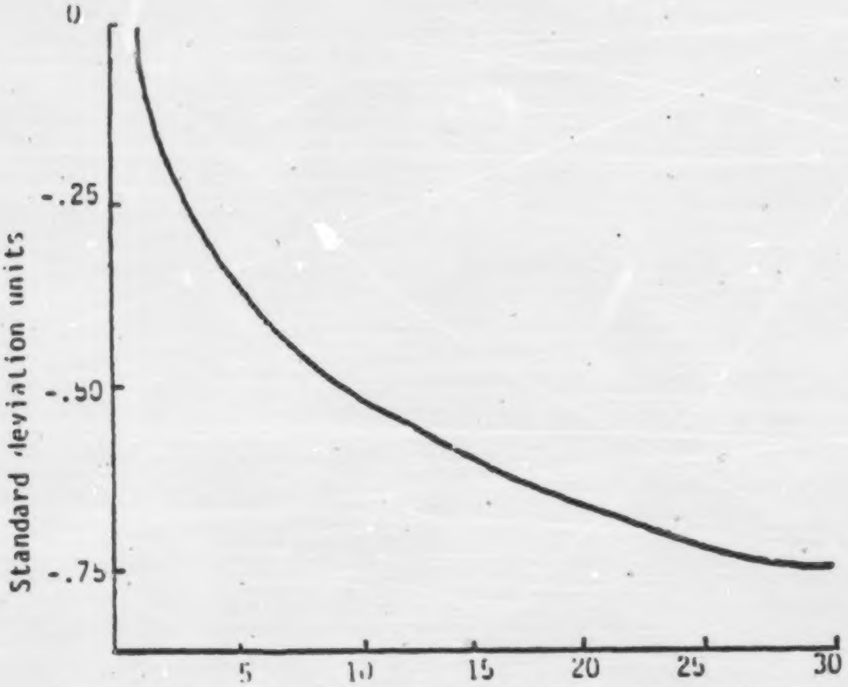


Figure 6.7. Data in Table 6.6 fitted to the log model.

An Alternative Log Model.

A model may have advantages if it avoids highly interdependent data sets created (as in the first model) by taking all pairwise differences in a study. Such an alternative model can be developed along the following lines.

Let $\overline{y}_c$ and s_c be the mean and standard deviation of the dependent variable for class-size $\underline{C}$ in one of $\underline{m}$ studies. For the $\underline{k}$ class-sizes in a particular study, order the groups from $C_1 < C_2 \ldots < C_k$. Arbitrarily set

$$\delta_k = 0 \text{ ; then}$$

$$\delta_{k-1} = \frac{\overline{y}_{k-1} - \overline{y}_k}{(s_{k-1} + s_k)/2} \text{ ,}$$

$$\delta_{k-2} = \delta_{k-1} + \frac{\overline{y}_{k-2} - \overline{y}_{k-1}}{(s_{k-2} + s_{k-1})/2} \text{ ,}$$

$$\delta_{k-3} = \delta_{k-2} + \frac{\overline{y}_{k-3} - \overline{y}_{k-2}}{(s_{k-3} + s_{k-2})/2} \text{ , and so on.} \qquad (12)$$

The data from the fourteen class-size experiments have been scaled via formula (12) and are recorded in Table 6.7.

The following model can be postulated for data of the form in (4);

$$\delta = -\beta \log_e C + (\alpha_1 D_1 + \ldots + \alpha_m D_m) + \varepsilon, \qquad (13)$$

The $\underline{\alpha} \cdot \underline{D}$ terms in (13) represent dummy variables and arbitrary level parameters for the $\underline{m}$ separate studies; $D_i = 1$ if a δ in question comes from the $\underline{i}$th study, and it equals zero otherwise. The parameters β and $(\alpha_1, \ldots, \alpha_m)$ can be estimated by regressing δ onto $\log_e C$. We have done so for the data in Table 6.7 and obtained a weighted least-squares estimate of β equal to 0.22. The estimates of the α's are unimportant. In this regression, each δ was weighted by k^{-1} so that each of the 14 studies would receive equal weight. The result is virtually identical to that obtained for the model in (10)

The model in (13) is more general and of more significance than the model in (10). Model (13) can be applied in a wide range of circumstances in which studies with quantitative independent variables are integrated. The first log term in (13) can be replaced by any mathematical function appropriate to a particular application. The important point about model (13) is that it simultaneously resolves the problems presented by different scales of measurement of $\underline{Y}$ and different values of $\underline{X}$ compared across studies.

Table 6.7

Data on the Relationship of Class-size and Achievement from Studies Using Random Assignment of Pupils.

Study Number	C, Size of Class	δ_c
1.	1.	.32
1.	25.	0
2.	1.	1.44
2.	3.	1.22
2.	25.	0
3.	17.	-.29
3.	35.	0
4.	28.	-.03
4.	112.	0
5.	1.	.89
5.	2.	.53
5.	5.	.31
5.	23.	0
6.	15.	.17
6.	30.	0
7.	16.	.09
7.	23.	.04
7.	30.	0
7.	37.	0
8.	20.	.15
8.	28.	0
9.	26.	.29
9.	50.	0
10.	1.	.65
10.	32.	0
11.	15.	1.05
11.	37.	.65
11.	60.	0
12.	1.	.30
12.	8.	0
13.	15.	.07
13.	45.	0
14.	1.	.95
14.	14.	.17
14.	30.	0

Non-Parametric Integration When the Independent Variable is Quantitative

The methods of the previous section assume a model for the relationship between the dependent and a quantitative independent variable. Standardized contrasts of the form $\Delta_{X_1-X_2}$ are used to estimate the parameters of the model. In many instances, too little will be known about the relationship to hypothesize even an approximate model. Then, perhaps, an approach modeled after Tukey's methods of exploratory data analysis might be more appropriate (Tukey, 1977). No functional relationship need be hypothesized, and the data themselves will determine the shape of the curve. An example will help clarify the approach, which may differ in details in particular applications.

Andrews, Guitar and Howie (1979) performed a meta-analysis of experimental studies of stuttering therapies. Effect sizes were calculated for 42 studies; all studies were pretest vs. posttest designs without control groups. Effects were assessed by comparing the post-test mean against the pretest mean and standardizing by the pretest standard deviation:

$$\Delta_{E-C} = \frac{\bar{y}_{post} - \bar{y}_{pre}}{s_{y_{pre}}} . \tag{14}$$

The 42 studies yielded 116 Δ's. These Δ's were categorized by the type of therapy applied, the duration of the therapy, type of outcome measure, and several other features of the therapy and the clients. Differences in average effect were obtained across types of therapy: Prolonged Speech therapy gave a $\bar{\Delta}_{E-C} = 1.65$ for 47 effects; at the other end of the scale, Systematic Desensitization gave a $\bar{\Delta}_{E-C} = 0.54$ for 5 effects (Andrews, Guitar & Howie, 1979, Table 3). No correlation was found between the number of months after therapy at which effects were measured and the size of effect. This lack of correlation seemed surprising and prompted the further search for a decay of effect across time that is reported below. The "follow-up time" variable and type of therapy are confounded in the Andrews stuttering data set. For example, Airflow therapy showed an average Δ of 0.92, but these outcomes were measured at 4.2 months after therapy on the average. On the other hand, Attitude therapy showed a $\overline{\Delta} = 0.85$ for an average follow-up time of 3.3 months. The only real difference between Attitude and Airflow average effects might be attributable to varying follow-up times for measurement of benefits. Likewise, the effect of different follow-up times may reflect therapy differences. For this reason, the pattern of decay in effects across time should be examined separately within each type of therapy. But another feature of the studies is also confounded with follow-up time and should be likewise controlled. Therapies differed with respect to the attention given to providing for post-therapy maintainence of the gains made during therapy. Andrews and his colleagues classified each study by whether there were _many_, _some_ or _no_ provisions made for maintainence of gains achieved during therapy. Thus, it seemed sensible to cross-classify effects by therapy type and maintainence provisions before examining the data for the decay of treatment phenomenon. Thus, 107 of the 116 effect sizes were cross-classified into the cells of an 8 x 3 (therapy type x maintainence provision) table, and the cell entries were averaged.

The averaging of effects resulted in an 8 x 3 table (see Table 6.8). The typical entry is a triplet of numbers of the form (a, b, c), where _a_ is the follow-up time in months, _b_ is the average $\overline{\Delta}_{0-a}$, and _c_ is the number of values averaged. Within a cell of Table 6.8 the entries were graphed in a connected line. Consider, for example, the cell for Rhythm therapy with many provisions for maintainence. The four data points can be graphed, as shown by the solid

Table 6.8

Follow-Up Time, Average Effect Size and Number of Effects Averaged Classifed by Type of Therapy and Provisions for Maintainence

	Maintainence Provisions		
Therapy Type	1: None	2: Some	3: Many
Airflow		1, .88, 1 3, .74, 1 16, .86, 1	
Rhythm	0, .66, 1 14, .76, 2		0, 1.26, 7 6, 1.57, 2 9, 1.60, 10 12, .86, 4
Shadow	0, .17, 1 14, .38, 1		
Gentle Onset	0, 1.12, 2 1, 1.38, 1 10, 1.12, 1 25, 1.15, 1		0, 2.37, 2 10, 1.52, 2
Biofeedback			0, .88, 2 12, 1.03, 2
Attitude	0, .71, 7 9, 1.11, 4		
Prolonged Speech	0, 2.02, 6 3, 2.42, 2 6, 1.27, 2 9, 2.17, 3 11, 1.77, 1		0, 1.62, 9 2, 2.02, 3 12, 1.16, 8 15, 1.16, 8 18, 1.36, 3
Desensitization	0, .69, 1 1, .89, 1 20, 1.07, 1	1, .01, 1 3, .03, 1	

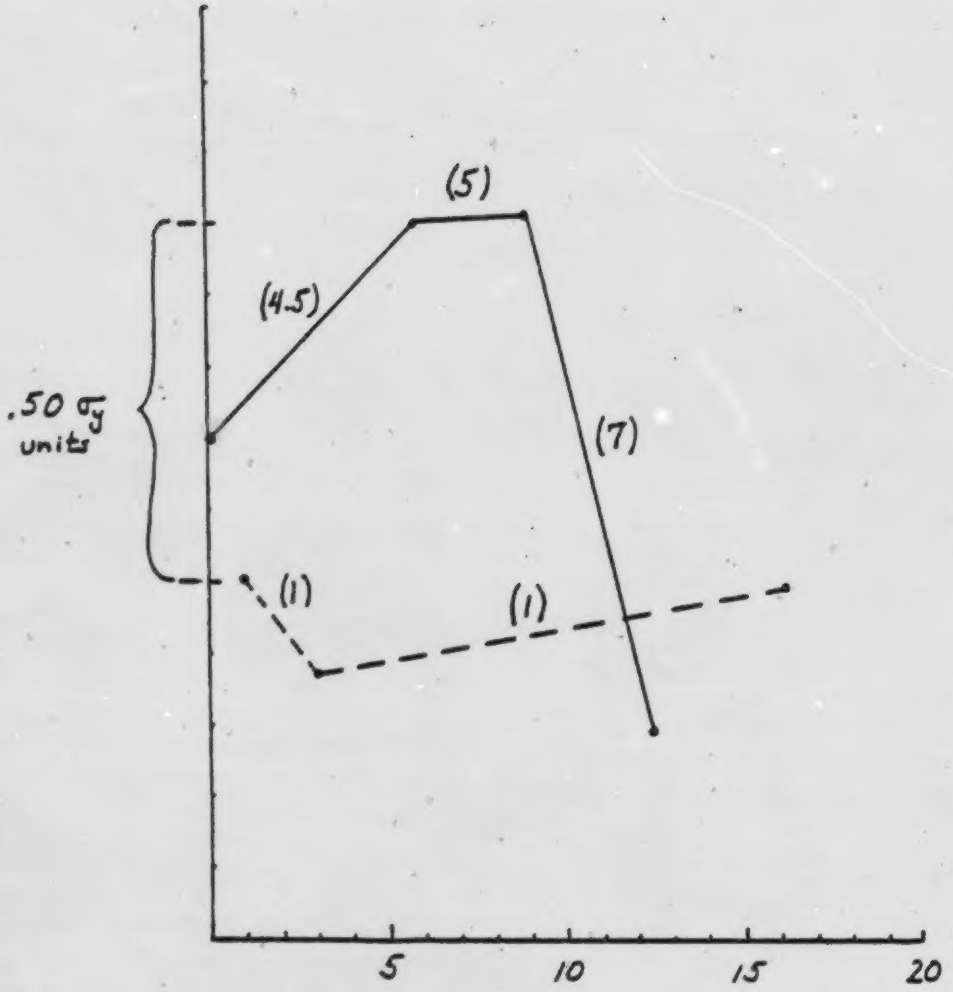


Figure 6.8. Graphs of effects over time for two cells of Table 6.8.

line in Figure 6.8. The broken line represents the three data points from Airflow therapy at the second maintainence level. The elevation of either line on the graph is immaterial; only the slope of the line relative to the abscissa is significant. The number in parentheses beside each line is the average of the number of effects, Δ_{o-a}, that exist at each end of the line; for example, the first segment of the solid line is based on 7 Δ's at zero months and 2 Δ's at six months--hence the weight $(7+2)/2 = 4.5$ for the line segment.

One approach to aggregating the data on slopes is to take a weighted average of all the lines above two successive months. For example, the slope of the solid line in Figure 4 between months 1 and 2 is $+.05 = \frac{1.57 - 1.26}{6 \text{ mos.}}$; the slope of the broken line is -.07. Since the weight for the solid line segment is 4.5 and for the dashed line, 1.0, the weighted average slope between months 1 and 2 is $[4.5(.05) + 1.0(-.07)]/(4.5 + 1.0) = + .028$.

If the above procedure were repeated for each successive pair of months and for all twelve lines that can be drawn from the data in Table 5, a complete aggregate curve is obtained. Such a curve is depicted in Figure 5. The curve shows a loss of benefits over the first twelve months after termination of therapy; the average loss is roughly one-half standard deviation. Although the general trend in the curve is unmistakably downward, not every intermediate twist and curve is to be taken seriously as a stable, replicable feature of the true relationship. Even though approximately twenty Δ's are still determining the slope of the aggregate curve in Figure 5 at 12 months post therapy, the estimates of the points on the curve are probably subject to a fairly large sampling error. Inferential techniques, perhaps drawing on Tukey's jackknife procedure (Mosteller and Tukey, 1968), would illuminate the question of the reliability of the determination of the curve.

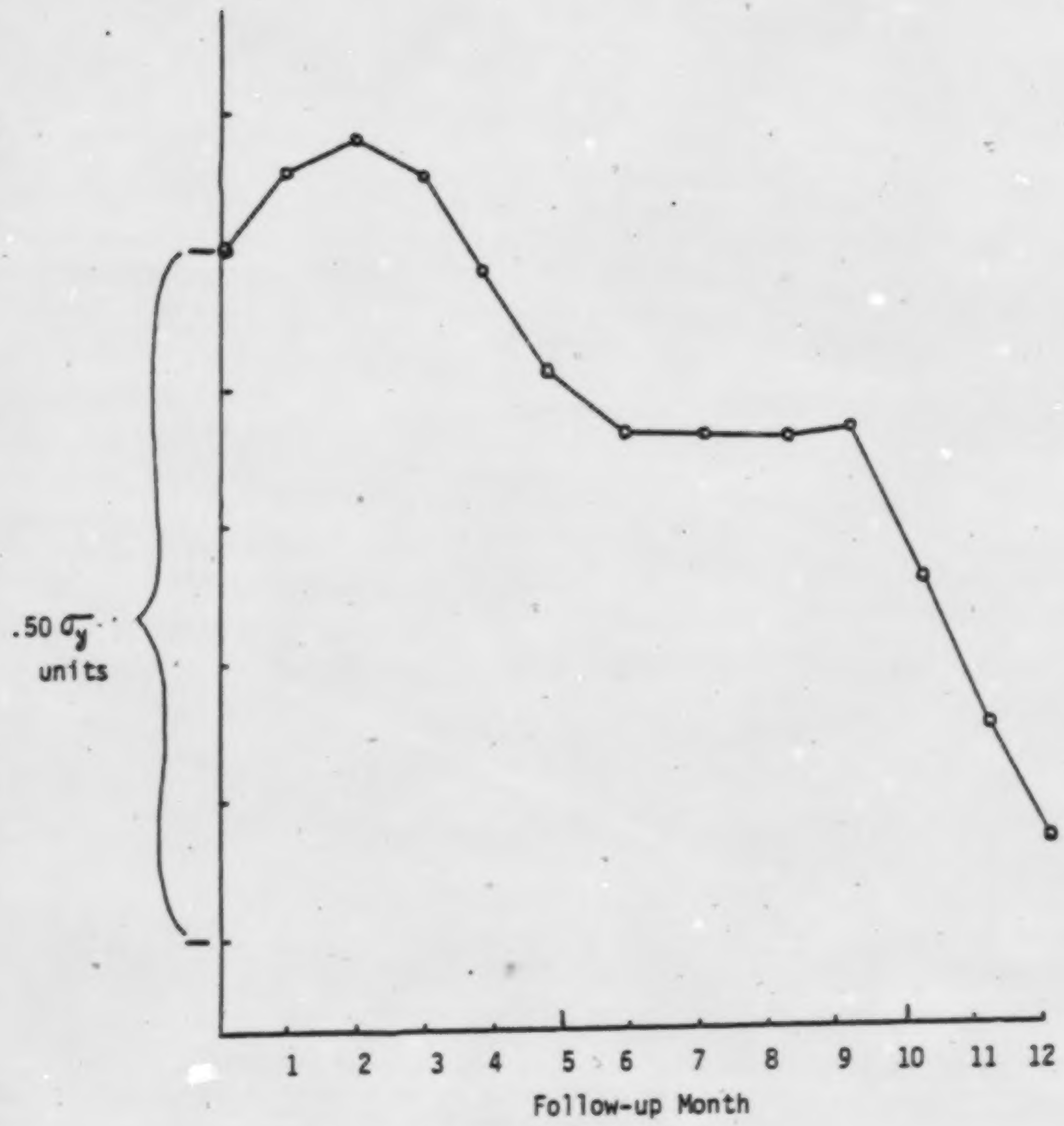


Figure 6.9. Aggregation by weighted averaging of data in Table 5 on the decay of stuttering therapy effects.

Aggregating Linear Slopes

An alternative approach was applied to the analysis of deterioration effects. This approach could be characterized as parametric to distinguish it from the non-parametric method illustrated above. Within each cell of Table 6.8. a straight trend line was fit to the $(t, \overline{\Delta}.)$ data by means of least-squares, i.e., the following model was fit by least-squares:

$\overline{\Delta}. = \beta_0 + \beta_1 t + \varepsilon$, where

$\overline{\Delta}$ is the average effect

t is the number of months after treatment that the dependent variable was measured.

These individual cell analyses number eleven. In each, estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ were obtained; in addition, the average number of Δ's for the data points in the cell was obtained. For example, for the cell "Airflow/Some Maintenance Provisions" in Table 6.8, the regression of $\overline{\Delta}.$ onto t for the three data points gives $\hat{\beta}_0 = .81025$ and $\hat{\beta}_1 = .00246$. In addition, since each $\overline{\Delta}.$ was based on $n = 1$, the average $\underline{n}$ is $\overline{n} = 1$. The regression equation spans the time interval 1 to 16 months, with a weight of $\overline{n} = 1.50$ and gives $\hat{\beta}_0 = .66000$ and $\hat{\beta}_1 = .00714$. In Table 6.9 appear the within cell regression lines, the follow-up interval spanned and the $\overline{n}$-weights.

The information in Table 6.9 can be integrated into a single curve by taking the $\overline{n}$-weighted average of all slopes, $\hat{\beta}_1$. Only those slopes are averaged at time point $t = t_i$ which were derived on data from a time interval that spans t_i. For example, the aggregate slope at $t = 0$ is a weighted average of all $\hat{\beta}_1$'s in Table 6.9 except those for "Airflow/Some" and "Desensitization/Some" which were based on intervals that begin at $t = 1$ month post-therapy.

Table 6.9

Within Cell Regression Lines, Time Interval and $\overline{n}$-weights for the Data in Table 6.8

Therapy/Maintenance Provision Combination	Regression of $\overline{\Delta}$ onto t:		Time Interval Spanned (in months)	$\overline{n}$-weight
	$\hat{\beta}_0$	$\hat{\beta}_1$		
Airflow/Some	.81025	.00246	1, 3, 16	1.00
Rhythm/None	.66000	.00714	0, 14	1.50
Rhythm/Many	1.45685	-.01990	0, 6, 9, 12	5.75
Shadow/None	.17000	.01500	0, 14	1.00
Gentle/None	1.22832	-.00398	0, 1, 10, 25	1.25
Gentle/Many	2.37000	-.08500	0, 10	2.00
Biofeedback/Many	.88000	.01250	0, 12	2.00
Attitude/None	.71000	.04444	0, 9	5.50
Prolonged Speech/None	2.08383	-.02652	0, 3, 6, 9, 11	2.80
Prolonged Speech/Many	1.79433	-.03514	0, 2, 12, 15, 18	6.20
Desensitization/None	.78026	.01472	0, 1, 20	1.00
Desensitization/Some	.00000	.01000	1, 3	1.00

Hence, for t = 0,

$$\hat{\beta}_1 = [1.50(.00714) + 5.75(-.01990) + \cdots + 1.00(.01472)]$$

$$\div (1.50 + 5.75 + \cdots + 1.00) = -.0094.$$

So the inclination of the curve at _t_ = 0 is .0094 units downward. At _t_ = 1, all twelve of the regression slopes in Table 6.9 are averaged because each of the regression lines was determined across a time span that included t = 1. The $\bar{n}$-weighted average is

$$\hat{\beta} = [1.00(.00246) + 1.50(.00714) + \cdots + 1.00(.01000)]$$

$$\div (1.00 + 1.50 + \cdots + 1.00) = -.0084.$$

In this manner, the aggregated slope of the curve is determined for each month from _t_ = 0 to _t_ = 17. The resulting aggregate curve is graphed along with the previously derived non-parametric curve in Figure 6.10.

In Figure 6.10, it is clear that the curve based on the weighted averaging of fitted straight lines is smoother and more regular than the non-parametric curve. This feature seems an advantage since the true curve of effects plotted against follow-up times probably wouldn't follow the jagged, irregular path of the non-parametric curve. But the aggregated curve based on linear slopes appears to have attenuated the size of the effect decay across time. For example, between 2 and 12 months, the non-parametric curves drops about .40 standard deviation units. Over the same interval, the curve from aggregated linear slopes drops only about .15 standard deviation units. This difference is so great as to cause one to search for a compromise solution.

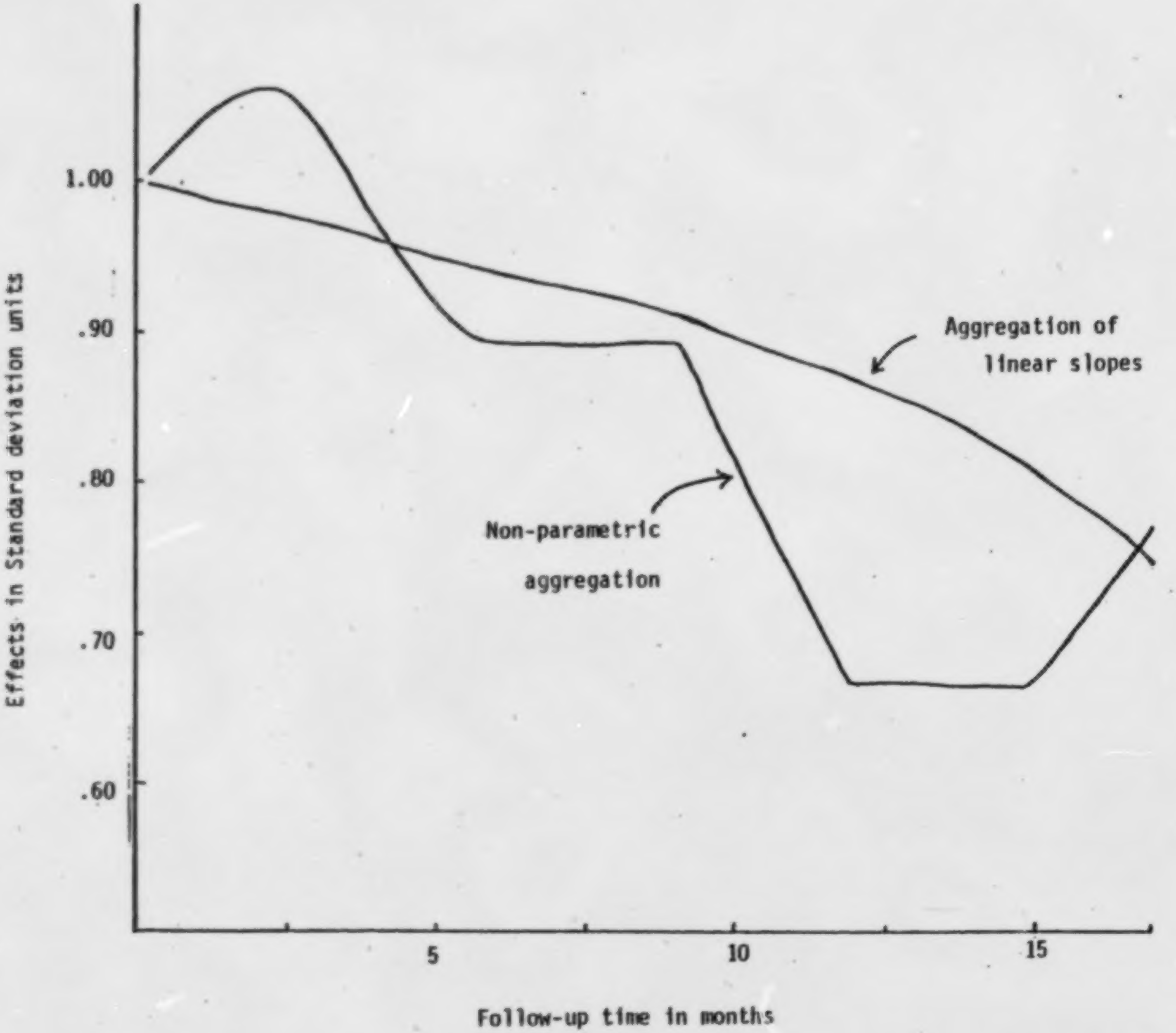


Figure 6.10. Comparison of non-parametric and linear methods of curve aggregation

Aggregating Quadratic Slopes

Fitting quadratic functions by least-squares estimation within each cell of Table 6.8 may produce a more satisfactory aggregate curve. Consider, for example, the cell "Airflow--'Some" Maintenance Provisions." The three pairs of points are as follows:

Follow-up time, t:	1	3	16
Effect size, Δ:	.88	.74	.86

These points can be fit to the quadratic equation

$$\bar{\Delta}. = \beta_0 + \beta_1 t + \beta_2 t^2 + e.$$

With three points and three parameters in the model, the fit of the euqation is perfect:

$$\hat{\bar{\Delta}}. = .9658 - .0911t + .0053\,t^2.$$

For example, at $t = 1$, the predicted effect is .88; at $t = 3$, $\hat{\Delta} = .74$; at $t = 16$, $\hat{\Delta} = .86$.

This single quadratic curve spans the time interval from 1 to 16 months. Its slope at any time t on the interval is given by the value of the derivative of the curve at the point t. In general, the slope of the curve at t, is given by

$$\text{Slope}(t_1) = \hat{\beta}_1 + 2\hat{\beta}t_1.$$

For example, the slope of the quadratic curve for "Airflow--'Some' Maintenance" at 2 months post-treatment is

$$\text{Slope}(t = 2) = \frac{d}{dt}(.9658 - .0911t + .0053t^2)\Big|_{t = 2}$$

$$= -.0911 + .0106t\Big|_{t = 2} = -.0699.$$

In words, then, the quadratic curve fit to the data has a slope of .07 standard deviation units downward at two months post-treatment.

This method of fitting quadratic curves can be applied to each cell of Table 6.8, provided that more than two follow-up times are present in a cell (at least three data points are required to estimate the three parameters of the quadratic curve). Consequently, six of the 12 non-empty cells in Table 6.8 must be eliminated. (An alternative approach not explored here would entail fitting straight lines in those cells with only two points and later aggregating their slopes with the slopes from the quadratic curves. This mixing of quadratic and straight line models is probably preferable to the elimination of two-data-point cells followed here.)

For each cell with sufficient data, a quadratic curve can be fitted via least-squares. Then the curve is differentiated to obtain the function describing the slope of the curve at any time $\underline{t}$. These slopes can be calculated for each value of $\underline{t}$ (to the nearest month, for example) across the time interval spanned by the data on which the curve was derived. Finally, for each value of $\underline{t}$ the slopes of the derived curves can be averaged, or averaged after some appropriate weighting, to form an aggregated curve. For the six quadratic curves fit to the data in Table 6.8, each slope was weighted by the average number of effect sizes in the cell (the same weight function applied in aggregating the data by the non-parametric and linear methods above).

The results of fitting the quadratic curves, the time span over which the curve stretches and the weight (average number of Δ's) for the six cells appear as Table 6.10. Suppose one wished to calculate the aggregate slope of the follow-up curve at $t = 16$ months post-treatment. From Table 6.10 it is seen that four cells contribute data to determining follow-up effects at 16 months: airflow-

Table 6.10

Quadratic Curves, Follow-up Time Spans and Weights (Average Number of Δ's) for the Data in Table 5

Cell	Time Span (in months)	Weight	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Airflow-Some	1 - 16	1.00	.9658	-.0911	.0053
Rhythm-Many	0 - 12	5.75	1.2413	.1741	-.0168
Gentle Onset-None	0 - 25	1.25	1.2471	-.0146	.0004
Prolonged Speech-None	0 - 11	2.80	2.1571	-.0818	.0050
Prolonged Speech-Many	0 - 18	6.20	1.8599	-.0857	.0029
Desensitization-None	0 - 20	1.00	0.6900	.2095	-.0095

some, gentle onset-none, prolonged speech-many, and desensitization-none. The first derivatives of the quadratic curves for these four cases and the weights associated with each curve are as follows:

	First Derivative	Weight
Airflow-some	$-.0911 + .0106t$	1.00
Gentle-onset-none	$-.0146 + .0008t$	1.25
Prolonged speech-many	$-.0857 + .0058t$	6.20
Desensitization-none	$.2095 - .0190t$	1.00

The aggregate slope at $t = 16$ is found by solving each first derivative at $t = 16$ and then forming the weighted average of the resulting four values:

$$\frac{1.00(.0785) + 1.25(-.0018) + 6.20(.0071) + 1.00(-.0945)}{1.00 + 1.25 + 6.20 + 1.00}$$

$$= \frac{+.0258}{9.45} = +.0027 .$$

Thus, the slope of the follow-up curve at 16 months is a rise of three-thousandths of a standard deviation per month--imperceptibly different from a horizontal line. In similar manner, the slopes of the quadratic curves in Table 6.10 were aggregated for each month from 0 to 17 and composite curve reflecting the proper slope at each month was drawn. This curve, referred to as the "aggregation of quadratic slopes" appears along with the non-parametric aggregated curve in Figure 6.11.

The aggregation of quadratic slopes clearly overcame the manifest shortcoming of the method of aggregating linear slopes, viz., the attenuation of effects. The quadratic curve is much more like the non-parametric curve than was the aggregation of linear slopes.

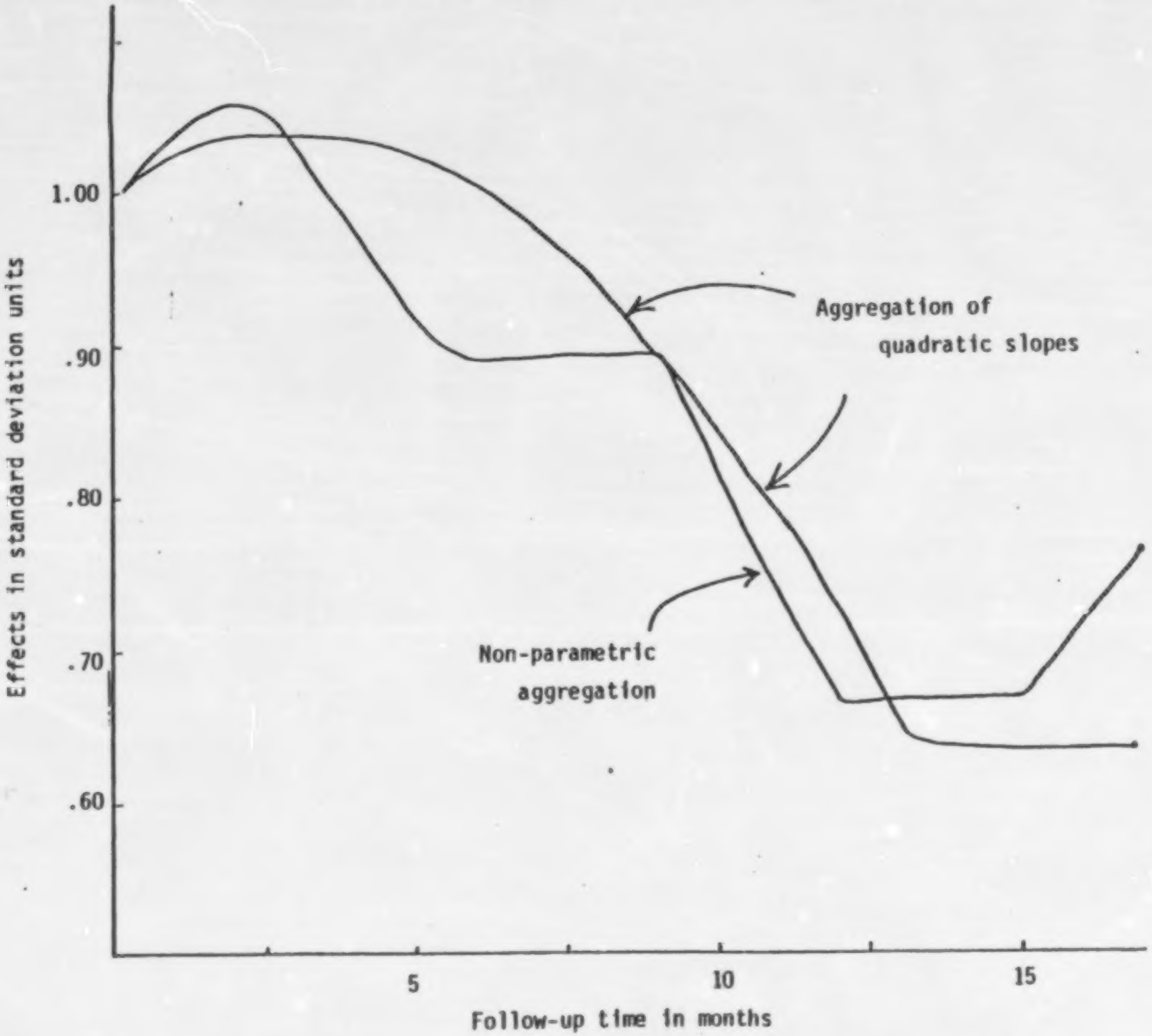


Figure 6.11. Comparison ofnon-parametric and quadratic methods of curve aggregation

INFERENTIAL METHODS OF META-ANALYSIS

The role of statistical inference in meta-analyses is somewhat controversial. Inference at the level of persons within studies (i.e., methods that treat persons as the unit of analysis) seems quite unnecessary; the rejection of hypotheses in such cases is nearly automatic and pro forma since even small integrative analyses encompasing twenty or so studies are likely to involve several hundred persons. The picture changes when one considers "studies" and the variability produced by their characteristics (e.g., location, date, investigator, types of subject, and the like). At this second level, one can readily imagine that even fifty or 100 studies may yield unstable findings, regardless of whether they subsume data from a thousand or many thousand persons. An investigator who subtly communicates his expectations of outcomes to his subjects affects all of them equally, and there is little comfort in there being 100 subjects or 1,000. So if any type of statistical inference ought to be undertaken in an integrative analysis, it should be carried out with "study" rather than "person" as the unit of analysis. But the prior question remains: should meta-analyses use inferential statistics?

The answer is, by no means, obvious. Inferential statistics seem to work well in two instances: randomized experiments and well-designed surveys with explicit sampling procedures. The classical theory of statistical inference assumes either the definition of a population and rigorous sampling from it or, as Fisher later showed, the randomization of units among conditions of an experiment. It works sensibly there; there is little doubt in these applications about what is meant when it is asserted that the confidence intervals cover the parameter with 95% probability or that the probability of the hypothesis being rejected incorrectly is 1%.

The typical integrative or meta-analysis seldom meets either condition of valid statistical inference. An attempt is made to locate every study on the topic being examined. Those studies that are located constitute a portion of a population of studies; but one hopes that the proportion is close to 100%, and one is under no illusions about the group of studies in hand being a random or probabilistic sample of the population. Rarely, a meta-analysis will be undertaken on a literature so large that it is impossible to read and analyze it all, even though one can describe, count and otherwise delineate the population of study. Then one might sensibly draw random (or stratified, cluster, two-stage random) samples of studies and apply classical inferential techniques with a legitimate warrant--as Miller (1978) was forced to do in his meta-analysis of the effects of psychoactive drugs.

The probability conclusions of inferential statistics depend on something like probabilistic sampling, or else they make no sense. There can be no question whether the relationship of a meta-analysis sample of studies to the population is similar to the experimental randomization upon which permutation test theory rests. It is not.

The arguments against inferential techniques in meta-analysis do not satisfy the appetite for some indication of the instability or unreliability of the results. When we showed our early work on psychotherapy (Smith and Glass, 1977) to John Tukey, he chided us for not presenting standard errors of the more important averages. Our recitation of the reasons for not broaching the inferential questions left him unconvinced; he felt that regardless of such complications, some rudimentary inferential calculations would be informative and useful. Since then, we have pursued inferential questions at the "study" level and through the application of Tukey and Mosteller's jackknife technique (an all-purpose approach to statistical inference for complex data sets where classical theory is lacking).

Whether the findings from a collection of studies are regarded as a sample from a hypothetical universe of studies, or they are in fact a sample from a well-defined population, problems of statistical inference arise. Significance tests or confidence intervals around estimates of averages or regression planes will indicate where the research literature is conclusive on a question and where the aggregated findings still leave doubts -- at least insofar as sampling error is concerned.

The inferential statistical problems of the meta-analysis of research are uniquely complex. The data set to be analyzed will invariably contain complicated patterns of statistical dependence. "Studies" cannot be considered the unit of data analysis without aggregating findings above the levels at which many interesting relationships can be studies. Each study is likely to yield more than one finding. An experiment comparing heterogeneous and homogeneous ability grouping might produce effect-size measures on three types of school achievement at four points in time; thus, 12 of the several hundred effect-size measures in an aggregate data set would have arisen from a single study. There is no simple answer to the question of how many independent units of information exist in the larger data set. One might attempt to impose some type of cluster or multiple-stage sampling framework on the data, but in the end this will probably restrict the movement of an imaginative data analyst. Two resolutions of the problem can be envisioned: one risky, the other complex.

The simple (but risky) solution is to regard each finding as independent of the others. The assumption is untrue, but practical. All inferential calculations could proceed on this independence assumption. The results (standard errors of means, of correlations, and of regression coefficients) could be reported with the qualification that they were calculated under the assumption of independence. This procedure might be useful because the effect of the dependence is almost surely to increase standard errors of estimates above what they would be if the same number of data points were independent. Thus, if 50 effect-size measures from 30 studies yielded an unsatisfactorily large standard error for the mean effect size, then it could be assumed safely that the standard error would be even larger if the complex dependence in the data were accounted for properly.

The matter of statistical efficiency and "lumpy" data can be described more formally by appealing to an analogy with cluster sampling in survey research. Imagine that "studies" are like clusters and effect size measures (or r's or any other appropriate description of findings) are like observations or cases within clusters. It is well-known from sampling theory (Cochran, 1963) that if m clusters each containing n elements are drawn randomly from a population in which the intra-cluster correlation of elements is denoted by ρ, then the variance error of the mean of the mn observations is given approximately by:

$$\mathrm{Var}^*(\overline{y}.) \doteq \frac{\sigma_y^2}{mn}\left[1 + (M - 1)\rho\right] , \qquad (15)$$

where σ_y^2 is the homogeneous within cluster variance of the observations.

The analogy with applications to meta-analysis can be drawn by associating studies with clusters and then ρ becomes the intra-study correlation of effect-sizes, say. It is instructive to notice in the above equation that intra-cluster (or "intra-study") correlation changes the variance of the mean, from what would be obtained under independence, by a factor of $1 + (m - 1)\rho$. It is improbable that ρ would ever be negative, hence the conclusion that intra-study correlation of findings in meta-analyses increases variance errors, thus decreasing the reliability of aggregates from what would be expected under independence.

Fortunately, the results from several extant meta-analyses can be used to investigate what a typical value of ρ might be. Then, the typical inflation of the variance error of the mean can be estimated. In Table 10 appear the intra-study correlation coefficients (of course, these are merely intra-class correlations) calculated from the data of seven meta-analyses.

Only one of the seven ρ's in Table 10 is below .50; they average .61, but they vary greatly about that average. Nonetheless, .60 gives a reasonably typical value of ρ with which to inquire further.

Under the assumption of independence of findings within studies, the variance error of an aggregate average of n findings within each of m studies is given by:

$$\mathrm{Var}(\bar{y}.) = \frac{\sigma^2}{mn} \, .$$

An intra-study correlation of findings increases the variance of the mean to:

$$\mathrm{Var}^*(\bar{y}.) = \frac{\sigma^2}{mn} \left[1 + (m - 1).6\right] .$$

The ratio of the latter to the former equals:

$$1 + (m - 1).6$$

Table 6.11

Intra-Study Correlation Coefficients from Seven Meta-Analyses

Investigator(s)	Topic	No. of Studies	No. of Findings Within Studies	ρ
Kavale ('79)	Psycholinguistic training	27	220	.24
Schlesinger, Mumford & Glass ('78)	Treatment of asthma	11	19	.85
Smith ('80)	Sex-bias in psychotherapy	34	60	.69
Glass *et al.* ('77)	Teacher indirectness & achievement	19	34	.90
Glass ('77)	Effects of psychotherapy on anxiety	26	39	.51
Smith, Glass & Miller ('80)	Psychotherapy	60	185	.60
Shavelson *et al.* ('77)	Stability of teacher effects	19	52	.50

which indicates the inflation of the variance error due to the non-independence of findings within studies. It is important to note that the inflation factor does not depend on the number of findings, _n_, within studies, but rather it depends on the number of studies, _m_.

Another way to view the inflation of the variance error of the mean due to non-independence is to express $\mathrm{Var}(\bar{y}.)$ as follows by dropping terms of order $1/m$:

$$\mathrm{Var}^*(\bar{y}.) = \frac{\sigma^2}{mn} + \frac{\rho}{n}\sigma^2 = \sigma^2\left(\frac{1}{mn} + \frac{\rho}{n}\right) . \tag{16}$$

This formulation shows that the variance of the mean is increased by $\sigma^2\rho/n$ due to the non-independence of findings within studies.

The following table illustrates the inflation of $\mathrm{Var}^*(\bar{y}.)$ over $\mathrm{Var}(\bar{y}.)$ because of non-independence. It is based on the typical intra-study correlation of .60 from Table 10 and an assumption of _n_ = 2 findings per study.

No. of Studies	_a_ $\mathrm{Var}(\bar{y}.)$	_b_ $\mathrm{Var}^*(\bar{y}.)$	b/a
5	$(.10)\sigma^2$	$(.34)\sigma^2$	3.4
10	$(.05)\sigma^2$	$(.32)\sigma^2$	6.4
20	$(.025)\sigma^2$	$(.31)\sigma^2$	12.4
50	$(.01)\sigma^2$	$(.304)\sigma^2$	30.4
100	$(.005)\sigma^2$	$(.302)\sigma^2$	60.4
500	$(.001)\sigma^2$	$(.3004)\sigma^2$	300.4

The calculations are remarkable. They show, for example, that given an intra-study clustering of .6 for 50 studies with two findings each, the variance error of the mean of all 100 findings is thirty times larger than the variance error one would suppose to be true assuming independence. Thus, statistical intuitions developed from experience with independent data sets must be held in check when dealing with the kinds of non-independence data typical of meta-analyses. Furthermore, it is important that statistical techniques applied to meta-analysis take account of the non-independent structure of the data, either by use of formulas for clustering such as illustrated here or by use of the jackknife technique.

Tukey's Jackknife

An inferential technique which takes account of the interdependencies in a large set of findings in a meta-analysis is Tukey's jackknife method (Mosteller & Tukey, 1968). Space does not permit a basic exposition of the jackknife technique. One suggestion and an example must suffice. In calculating the "pseudovalues" in the jackknife method, some portion of the data set is discarded, and the sample estimate of the parameter of interest is calculated. In a meta-analysis, the portion of data eliminated should correspond to all those findings (e.g., effect sizes or correlation coefficients) arising from a particular study. Thus there will be as many pseudovalues as there are studies. The method will be illustrated on a small portion of the data from a meta-analysis of psychotherapy outcome studies.

The data in Table 6.12 represent 39 effect-size measures from 26 experimental studies in which behavioral and nonbehavioral psychotherapies were compared for their effects on fear and anxiety. The effect-size measure was defined as $\Delta = (\bar{X}_{beh.} - \bar{X}_{nonbeh.})/S_X$. For example, study 1 produced two measures of experimental effect, the first of which shows the nonbehavioral therapy as slightly superior to the behavioral therapy, and the second of which shows the behavioral therapy nearly three-fourths of a standard deviation superior to the nonbehavioral therapy. The first step in establishing a jackknife confidence interval on the mean effect size is to average the 39 effect-size measures to obtain $\overline{X}$. Second, 26 partial means, $\overline{X}_{-i}$, are calculated by eliminating each study in turn; for example, the first partial mean is based on the 37 effect-size measures remaining after the effect sizes from study 1 (-.10, .74) are removed. Third, 26 pseudovalues are calculated as follows: $\hat{\theta}_i = 26\overline{X}. - 25\overline{X}_{-i}$. The 26 pseudovalues can safely be regarded as sample of observations of normally distributed independent variables, with expected value approximately equal to the true mean effect size and variance $\sigma^2_{\hat{\theta}}$. Thus, the set of pseudo values, $\hat{\theta}_i$, can be treated as an ordinary sample of data to which t- distribution methods can be applied. The right-hand side of Table 6.12 lists the calculations for the 95 percent confidence interval on the true effect size; the interval does not quite span zero, indicating a statistically reliable superiority of the behavioral therapies. By comparison, a t-method 95 percent confidence interval on the population mean effect size calculated from the 39 effect-size measures, assuming independent observations, extends from -.10 to + .50.

Statistical inferential methods on the type of data illustrated here could play a role in directing future research. From standard errors of averages and confidence regions around regression planes, one can determine where parameters are sharply estimated by the current body of research studies and where sample estimates remain poor. The simple cross-tabluation of the characteristics of studies completed is helpful for the same purpose. However, it must be pointed out that the number of studies needed to estimate accurately an aggregate effect size is partly a function of the variance of effect sizes. For example, 5 studies may determine accurately the effect of amphetamines on hyperactive 8-year-olds, whereas 20 studies may be needed to achieve the same accuracy with 12-year-olds if the effects are fundamentally more variable for older children.

Table 6.12

Illustration of Application of the Jackknife Technique of Interval Estimation of Mean Effect Size

Study No.	Effect-Size Measures	Pseudo Values $\hat{\theta}_i = 26\bar{X}_. - 25\bar{X}_i$	Calculations
1	− .10		
....	.74	.366	$N = 39$ effect-size measures
2	.43		$n = 26$ studies
....	.45	.528	
3	.65	.493	$\bar{X}_{\hat{\theta}} = .186$
4	.52	.407	
5	.20	.197	$s_{\hat{\theta}} = .457$
6	− .16	−.040	
7	− .50	−.264	
8	3.35		
....	−2.82	.291	95% jackknife confidence
9	.18	.184	interval on μ:
10	.51	.278	
11	− .39	−.191	$_{.975}t_{25} = 2.06$
12	− .95	−.560	
13	.33	.282	
14	.12	.144	$\bar{X}_{\hat{\theta}} \pm ts_{\hat{\theta}}/\sqrt{n} =$
15	.08	.118	$.186 \pm (2.06)(.457)/\sqrt{26} = (.002, .371)$
16	1.90	1.315	
17	− .44	−.224	
18	−1.00	−.593	
19	.06		
....	.20		
....	.10		
....	.00	−.097	
20	.64	.486	
21	.59		
....	.96	.980	
22	.05		
....	.20	.102	
23	.01	.072	
24	.12		
....	.00		
....	.14		
....	− .28	−.368	
25	− .22	−.079	
26	1.28		
....	.24		
....	.24	1.016	

The illustration above showed that a confidence interval based on jackknifing on "study" as the unit of analysis was narrower than the confidence interval calculated by traditional methods with individual Δ's as the unit of analysis. This was unexpected and contrary to the illustration to be presented here. It probably is due to the fact that the largest positive and largest negative values of Δ arose from the same study. A recent application of the jackknife to meta-analysis by Haertel, Walberg and Haertel (1979) gave results more in accord with expectations. When multiple linear regression weights were jackknifed using "study" as the unit, the _t_--statistics for the significance of the differences of the beta-weights from zero were nearly always smaller for the jackknife estimates than for the conventional estimates (Table 4 of Haertel, Walberg and Haertel, 1979).

An illustration will indicate the lines along which the jackknife approach to statistical inference in meta-analysis can be applied. The class-size and achievement analysis above can serve as the illustration. A total of 108 comparisons of achievement in smaller and larger classes was available to fit the logarithmic curve. These 108 comparisons actually arose from 14 different studies. The multiplicity of data arose both from multiple comparisons with a study (a study comparing four class sizes produced six Δ's) and multiple achievement measures for individual comparisons. (The complete data set appears in Glass and Smith, 1978.) A traditional inferential analysis that takes no regard of the complex interdependencies of the data set (108 Δ 's corresponding to only 30 unique comparisons of class-size arising from only 14 studies) would proceed along the following lines.

The least-squares regression of Δ_{S-L} onto $\log_e(L/S)$ has the solution:

$$\hat{\beta} = \frac{\Sigma\Delta_{S-L}(\log_e L/S)}{\Sigma(\log_e L/S)^2} \tag{17}$$

For the 108 data points,

$$\hat{\beta} = \frac{108.780}{385.745} = .2820 .$$

The estimate of residual variance equals:

$$\hat{\sigma}_e^2 = .1823 .$$

From traditional least-squares theory, it can be shown that:

$$\sigma_{\hat{\beta}}^2 = \sigma_e^2 \left[\Sigma(\log_e L/S)^2\right]^{-1}$$

Thus, in the example,

$$\hat{\sigma}_{\hat{\beta}} = \sqrt{.1823(385.745)^{-1}} = .02174.$$

Assuming normal distributions of estimates of β, the 95% confidence interval on β is given by:

$$\hat{\beta} \pm 1.98\ \hat{\sigma}_{\hat{\beta}} = .2820 \pm .0430 = (.2390,\ .3250).$$

The results of the interval estimation prove to be quite different when the jackknife method is used to take account of variation at the study level. The first step in calculating the jackknife interval on β involves the calculation of all 14 pseudo-values, one for each study, by the formula:

$$\hat{\theta}_{-i} = 14\hat{\beta} - 13\hat{\beta}_{-i}\text{ , where}$$

$\hat{\beta}_{-i}$ is the estimate of β calculated by excluding all pairs of Δ_{S-L} and $\log_e L/S$ that arise from the *i*th study.

Using the earlier calculations on the entire data set, it can be computed that:

$$\hat{\theta}_{-i} = 3.948 - \frac{108.780 - \sum_{1}^{ni} \Delta \log_e(L/S)}{385.745 - \sum_{1}^{ni} (\log_e L/S)^2} \quad ,$$

where the summation is over all pairs of values of Δ and $\log_e L/S$ that appear in the *i*th study.

The fourteen values of $\hat{\beta}_{-i}$ for the data appear below coded by the study number used in Glass and Smith (1978):

Study No.	$\hat{\beta}_{-i}$	$\hat{\theta}_{-i}$
001	.28611	.222057
003	.216408	1.134696
006	.284079	.254973
008	.285092	.241804
009	.283260	.265620
016	.282092	.280810
035	.286599	.222213
049	.281716	.285692
052	.281494	.288578
055	.312188	-.110444
058	.277897	.335339
061	.281980	.282226
073	.282685	.273095
077	.293232	.135984

$\bar{\hat{\theta}}_{.} = .293760,$

$s_{\hat{\theta}} = .265047$

The 95% confidence interval on β is now calculated by the formula:

$$\bar{\bar{\theta}}_{\cdot} \pm {}_{.975}t_{df}\,\hat{\sigma}_{\hat{\theta}} \;/\sqrt{n} \quad ,$$

where n is the number of studies and df is n - 1, in this case, but not generally.

For the data of this illustration, the above formula takes the value:

$$\begin{aligned} &.293760 \pm 2.14\ (.265047)/\sqrt{14} \\ &= .293760 \pm .151590 \\ &= (.1422,\ .4454). \end{aligned}$$

This jackknife interval on β is more than 350% wide than the interval calculated earlier by conventional methods that treated each pair of values Δ and $\log_e L/S$ as an independent data point. The jackknife methods appears to be appropriate and equal to the task of handling data sets interlaced with complicated dependencies.

Generalized Least-Squares

The methods illustrated on the class-size data above are ordinary least-squares analysis (OLS) and Jackknife (JK) analysis. There exists a third means of analysis that is theoretically more rigorous and may prove superior to the putatively inappropriate OLS and the unknown JK analysis. The third method is the method of generalized least-squares analysis (GLS).

OLS is the traditional method of linear estimation based on a model of independently and normally distributed errors. It is, in fact, a special case of the method of GLS, which permits the errors in the linear model to be correlated. Correlated errors prevail in the type of data that are fitted to the logarithmic model in meta-analyses.

Suppose, to begin with a simple example, that a study of the relationship between class-size and achievement is performed where achievement is compared among class-sizes of n_1, n_2 and n_3 pupils (assume the n's increase in size from n_1, to n_3). From the logarithmic model,

$$z_{ij} = \beta \log n_j + e_{ij}$$

for the ith pupil in the jth class. It is assumed that e_{ij} are independently and normally distributed with variance σ^2. In order to remove arbitrary scale factors and fit the model, the class means must be paired, differenced and standardized to form delta measures; e.g.,

$$\Delta_{n_1-n_2} = \beta \log \left(\frac{n_1}{n_2}\right) + (\bar{e}_{\cdot 1} - \bar{e}_{\cdot 2}).$$

Now, the random variable Δ has a normal distirbution with

$$\text{mean} = \beta \log(n_1/n_2), \text{ and}$$

$$\text{variance} = \text{Var}(\bar{e}_{\cdot 1} + \bar{e}_{\cdot 2}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}.$$

There are three possible pairs of the class-sizes n_1, n_2 and n_3; thus there are three possible Δ's. However, the deltas are constrained by the restriction that

$$\Delta_{n_1-n_3} = \Delta_{n_1-n_2} + \Delta_{n_2-n_3}.$$

Thus, one of the three adds no information to the remaining two; only two deltas need be considered. (In the more general case of J class-sizes, there are $J(J-1)/2$ possible deltas, but only $J-1$ of these are free to vary.) Thus, the available information is completely contained in any nonredundant subset of $J-1$ deltas. It will be convenient to work with only those deltas that are formed by comparing each class-size in turn with the smallest class-size, e.g.,

$$\Delta_{n_1-n_j} \text{ where } n_j > n_1.$$

In the three class-size comparison, the deltas will be

$$\Delta_{n_1-n_2} \text{ and } \Delta_{n_1-n_3}.$$

We have already seen that $\Delta_{n_1-n_2}$ has error variance equal to $\sigma^2(1/n_1 + 1/n_2)$. Likewise, $\Delta_{n_1-n_3}$ has error variance equal to $\sigma^2(1/n_1 + 1/n_3)$. It remains to determine the covariance of these two deltas:

$$\begin{aligned}
&\text{Covar}\,(\Delta_{n_1-n_2}, \Delta_{n_1-n_3}) = \\
&\text{Covar}\,(\bar{e}_{\cdot 1} - \bar{e}_{\cdot 2}, \bar{e}_{\cdot 1} - \bar{e}_{\cdot 3}) = \\
&\text{Covar}\,(\bar{e}_{\cdot 1}\bar{e}_{\cdot 1}) - 0 - 0 + 0 = \\
&\text{Var}(\bar{e}_{\cdot 1}) = \sigma^2/n_1.
\end{aligned}$$

It should be clear that in a set of J-1 deltas formed by comparing each n_j in turn with n_1, that each delta has variance given by

$$\text{Var}(\Delta_{n_1-n_j}) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_j}\right),$$

and each pair of deltas has covariance given by

$$\text{Covar}(\Delta_{n_1-n_j}, \Delta_{n_1-n_{j*}}) = \sigma^2/n_1.$$

Hence, the set of two deltas in our example has the following variance-covariance matrix of errors:

$$\Sigma_\Delta = \sigma^2 \begin{bmatrix} \frac{1}{n_1} + \frac{1}{n_2} & \frac{1}{n_1} \\ \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_3} \end{bmatrix} \tag{18}$$

A general linear model could now be stated for the two deltas:

$$\Delta_{n_1 - n_j} = \beta \log(n_1/n_j) + \varepsilon$$

where the vector of ε's are distributed normally with zero mean vector and variance-covariance matrix in formula 18 above.

Denoting the variance-covariance matrix of errors by Σ_ε, then Johnston (1972) shows that the generalized least-squares solution for β is contained in the following quantities:

$$\hat{\beta} = (X^T \Sigma_\varepsilon^{-1} X)^{-1} X^T \Sigma_\varepsilon^{-1} \Delta, \tag{19}$$

where Δ is the vector of deltas (two, in the example), and X is the matrix of independent variable values (in the example, a 2 x 1 vector with entries $\log(n_1/n_2)$ and $\log(n_1/n_3)$),

$$\mathrm{Var}(\hat{\beta}) = \sigma_\varepsilon^2 (X^T \Sigma_\varepsilon^{-1} X)^{-1} \tag{20}$$

and an unbiased estimate of σ^2 is given by

$$\hat{\sigma}_\varepsilon^2 = (\Delta - X\hat{\beta})^T \Sigma_\varepsilon^{-1} (\Delta - X\hat{\beta})/(N - k), \tag{21}$$

where N is the number of deltas and k is the number of parameters estimated (one, in the example).

In a typical meta-analysis, deltas will arise from more than one study. Thus, there may be two deltas from Study #1 ($J=3$) and three deltas from Study #2 ($J=4$). This arrangement of data does not substantially complicate the GLS analysis outlined above. The vector of deltas is now of order 5 x 1 and the variance-covariance matrix of errors, ε, is a block-diagonal matrix of order 5 x 5:

$$\Sigma_\varepsilon = \sigma^2 \begin{bmatrix} \frac{1}{n_1} + \frac{1}{n_2} & \frac{1}{n_1} & 0 & 0 & 0 \\ \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_3} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{n_1} + \frac{1}{n_2} & \frac{1}{n_1} & \frac{1}{n_1} \\ 0 & 0 & \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_3} & \frac{1}{n_1} \\ 0 & 0 & \frac{1}{n_1} & \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_4} \end{bmatrix},$$

where the n_1 in Study # 1 may be different from the n_1 in Study #2 (and likewise for n_2, . . .).

The block diagonal matrix Σ_ε in formulas (19), (20) and (21) yields the proper estimate of β, its standard error, and an estimate of error variance. The distribution of β divided by its estimated standard error is known to be Student's t-distribution with degrees of freedom equal to N-1, where N is the number of deltas (there being J-1 deltas for each study) (Johnston, 1972, p. 210).

The above argument appears to be mathematically complete and appropriate to the inferential problems of fitting and testing the logarithmic model. A Monte Carlo study is not strictly required--failing the discovery of some flaws in the mathematics--but it will be useful to check the validity of the GLS procedure while carrying out a Monte Carlo study to check the usefulness of the OLS and JK solutions. One knows _a priori_ that the OLS and JK confidence intervals do not have complete mathematical justifications; the OLS intervals are likely to be uselessly inexact and, as always, the accuracy of the approximation upon which the JK intervals are based must be checked.

In the following section, the results of a Monte Carlo simulation are presented in which the accuracy of confidence intervals constructed by the GLS, OLS and JK methods is compared.

Monte Carlo Study

A Monte Carlo study was conducted to check the validity of OLS, GLS and JK confidence intervals. The structure of the simulation (i.e., number of studies, number of class-sizes compared, and the sizes of classes compared) was chosen to duplicate exactly the data set in the meta-analysis of class-size and achievement (Glass and Smith, 1979). The data set structure is as follows:

Table 6.13

Structure of the Data Set Used in the Monte Carlo Study

Study No.	Class Sizes n_1	n_2	n_3	n_4 ....
1	1	3	25	
2	1	25		
3	17	35		
4	28	112		
5	1	2	5	23
6	15	30		
7	16	23	30	37
8	20	28		
9	26	50		
10	1	32		
11	15	37	60	
12	1	8		
13	15	45		
14	1	14	30	

For example, in Study #1, three class-sizes were compared: 1, 3 and 25. This study gives rise to two values of delta: Δ_{1-3} and Δ_{1-25}. In Study #4, two class-sizes are compared yielding a single delta: Δ_{28-112}.

Given the above data structure, there are only two parameters of the logarithmic model that need to be specified: the value of β and the error variance σ_e^2 (N.B.: this error variance describes error in observations of individuals; it is not the same as the error ε). The value of β can be specified without restriction; in the simulations, values of .25, .50 and 1.00 were used. The error variance, σ_e^2, is specified in a round-about way by first specifying a value for the linear correlation between z and $\log(n_1/n_2)$ in the model

$$z = \beta \log(n_1/n_2) + e, \text{ and}$$

then solving for σ_e^2 assuming that z has unit variance. In the simulations reported here, the linear correlation, ρ, between z and $\log(n_1/n_2)$ was taken to be either .65 or .85. Hence, the corresponding error variances equal

$$\sigma_e^2 = \sqrt{1 - .65^2} = 0.76;$$

$$\sigma_e^2 = \sqrt{1 - .85^2} = 0.53.$$

The steps in the simulation proceeded as follows:

Step 1. Having specified values of n_1, n_2, β and ρ (say, $\beta = .5$, $\rho = .85$), scores are generated according to the model

$$z = \beta \log(n_1/n_2) + e.$$

Step 2. Deltas are calculated via

$$\Delta_{n_1-n_2} = \frac{\bar{z}_1 - \bar{z}_2}{\sigma_z} .$$

Step 3. In this way, all the deltas in the 14-study data set specified above are calculated.

Step 4. The ordinary least-squares (OLS) estimate of β is calculated in the usual fashion from the 30 deltas that arise from Table 6.13. The $1 - \alpha$ confidence interval on β is calculated from

$$\hat{\beta} \pm {}_{1-\alpha/2}t_{29}\, \hat{\sigma}_{\hat{\beta}} .$$

Step 5. The jackknife (JK) confidence interval on β is calculated by means of the 14 pseudo-values arising from the data structure in Table 6.13 and then by means of the formula

$$\bar{\bar{\theta}}. \pm {}_{1-\alpha/2}t_{13}\, \hat{\sigma}_{\hat{\theta}}/\sqrt{14} .$$

Step 6. The generalized least-squares (GLS) confidence interval on β is calculated via

$$\hat{\beta} \pm {}_{1-\alpha/2}t_{20}\, \hat{\sigma}_{\hat{\beta}} \quad \text{where}$$

the estimates are given in formulas (19), (20) and (21) above.

Step 7. Each of the three intervals was calculated for each single simulation and it was recorded whether the 90 percent, 95 percent and 99 percent confidence intervals captured the true value of β. The simulation was repeated 1,000 times and the proportions of intervals capturing the parameter were counted.

The results appear in the following table for the 90 percent confidence coefficient. The results for the 95 percent and 99 percent confidence intervals appear in Tables 6.15 and 6.16 .

Table 6.14

Empirical Confidence Coefficients for True $\alpha = .10$ for Ordinary Least-squares (OLS), Jackknife (JK) and Generalized Least-squares (GLS) Confidence Intervals

ρ	β	Method of Estimation	Empirical Confidence Coefficient
.65	.25	OLS	.678
		JK	.857
		GLS	.900
	.50	OLS	.646
		JK	.845
		GLS	.909
	1.00	OLS	.641
		JK	.857
		GLS	.910
.85	.25	OLS	.653
		JK	.860
		GLS	.894
	.50	OLS	.647
		JK	.852
		GLS	.906
	1.00	OLS	.642
		JK	.854
		GLS	.897

Table 6.15

Empirical Confidence Coefficients for True $\alpha = .05$ for Ordinary Least-squares (OLS), Jackknife (JK) and Generalized Least-squares (GLS) Confidence Intervals

ρ	β	Method of Estimation	Empirical Confidence Coefficient
.65	.25	OLS	.786
		JK	.917
		GLS	.955
	.50	OLS	.740
		JK	.905
		GLS	.949
	1.00	OLS	.744
		JK	.912
		GLS	.956
.85	.25	OLS	.742
		JK	.914
		GLS	.947
	.50	OLS	.771
		JK	.929
		GLS	.947
	1.00	OLS	.734
		JK	.914
		GLS	.943

Table 6.16

Empirical Confidence Coefficients for True α =.01 for Ordinary Least-squares (OLS), Jackknife (JK) and Generalized Least-squares (GLS) Confidence Intervals

ρ	β	Method of Estimation	Empirical Confidence Coefficient
.65	.25	OLS	.866
		JK	.966
		GLS	.993
	.50	OLS	.871
		JK	.973
		GLS	.987
	1.00	OLS	.876
		JK	.969
		GLS	.988
.85	.25	OLS	.879
		JK	.973
		GLS	.994
	.50	OLS	.876
		JK	.981
		GLS	.991
	1.00	OLS	.869
		JK	.968
		GLS	.983

The results in Tables 6.14 and 6.16 are remarkably similar and the findings are clear. The GLS method is accurate; it yields the confidence coefficient that one expects to have when referencing the 1 - $\alpha/2$ percentiles of the proper t-distribution. The empirical and theoretical confidence coefficients were never more than .01 units discrepant--a discrepancy well within the bounds of sampling error for 1,000 cases, as it must be since the GLS solution is mathematically correct. By comparison, the OLS confidence intervals were grossly in error. For example with $\beta = 1.0$ and $\rho = .85$, the nominal 90 percent OLS confidence interval around $\hat{\beta}$ has only .642 probability of capturing the parameter value of 1.0--an error in the expected confidence coefficient of roughly one-third.

The JK confidence coefficients are more accurate than the OLS coefficients but they are probably unacceptably discrepant from theoretical values, in absolute terms, and they are clearly less accurate than the GLS confidence intervals. For example, for $\beta = .50$ and $\rho = .65$, the nominal 90 percent JK interval has actual confidence coefficient of 84.5%, an error of over 5 percentage points, whereas the GLS interval, as expected, shows an empirical confidence coefficient equal (within sampling error) to the theoretical value.

A Monte Carlo simulation showed the generalized least-squares confidence intervals on β of the logarithmic model to be accurate, according to theory. The ordinary least-squares confidence intervals proved to be grossly inaccurate and unacceptable--victims of the non-independence of the Δ's from which the logarithmic model is fitted. The jackknife confidence intervals (although not as inaccurate as the OLS intervals and although possibly capable of being improved by proper normalizing transformations yet to be discovered) were less accurate than the GLS intervals.

The method of generalized least-squares is an accurate method of interval estimation of β in the logarithmic model which finds frequent application in problems of meta-analysis.

CHAPTER SEVEN

AN EVALUATION OF META-ANALYSIS

The approach to research integration referred to as "meta-analysis" is nothing more than the attitude of data analysis applied to quantitative summaries of individual experiments. By recording the properties of studies and their findings in quantitative terms, the meta-analysis of research invites one who would integrate numerous and diverse findings to apply the full power of statistical methods to the task. Thus it is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis.

A tenet of evaluation theory is that self-assessment is alwyas more suspect than assessment by a neutral party. There is a tone of false promise in professing to criticize an endeavor in which one has invested himself heavily. Although we cannot promise to deal with the strengths and weaknesses of the meta-analysis approach with an even hand, we can assure the reader that most of the objections raised against the procedure by critics of earlier applications are recorded and discussed below. Applications of meta-analysis to research in psychotherapy, school class-size, special education and other problems have produced many technical criticisms. Among the persons commenting on meta-analysis are the following: Mansfield & Busse (1977), Bandura (1987), Eysenck (1978a), Gallo (1978), Jackson (1978), Paul (1978), Presby (1978), Walberg (1978), Anonymous (1979), Gillen (1979), Rimland (1979), Simpson (1980), Eysenck (1978b), Shapiro (1977), Cook and Leviton (1980), Hunter (1979), Roid, Brodsky and Bigelow (1979).

1. The Apples and Oranges Problem

   It is illogical to compare "different" studies, i.e., studies done with different measuring techniques, different types of persons, and the like.

2. Use of Data From "Poor" Studies

   Meta-analysis advocates low standards of quality for research. It accepts uncritically the findings from studies that are poorly designed or are otherwise of low quality. Aggregated conclusions should only be based on the findings of "good" studies.

3. Selection Bias in Reported Research

   Meta-analysis is dependent on the findings that researchers report. Its findings will be biased if, as is surely true, there are systematic differences among the results of research that appear in journals vs. books vs. theses vs. unpublished papers.

4. Lumpy (Non-Independent) Data

   Meta-analyses are conducted on large data sets in which multiple results are derived from the same study; this renders the data non-independent and gives one a mistaken impression of the reliability of the results.

In the remainder of this section, these criticisms will be addressed with counterarguments and data accumulated from several extant meta-analyses.

Criticism #1 - The Apples and Oranges Problem. The meta-analysis approach to research integration mixes apples and oranges. It makes no sense to integrate the findings of different studies.

The worry is often encountered that in combining or integrating studies, one is forcing incommensurable studies together, or trying to make different studies answer the same question, or "mixing apples and oranges." Implicit in this concern is the belief that only studies that are the same in certain respects can be aggregated. "A study's dependent variables and those independent variables which are measured must be measured in the same way as, or in a way subject to a conversion into, those employed in the rest of the studies" (Light and Smith, 1971, p. 449). This thesis should be clarified in at least two ways: "Same" is not defined and the respects in which comparable studies must be the same are unspecified. The claim that only studies which are the same in all respects can be compared is self-contradictory; there is no need to compare them since they would obviously have the same findings within statistical error. The only studies which need to be compared or integrated are different studies. Yet it is intuitively clear some differences among studies are so large or critical that no one is interested in their integration. What, for example, is to be made of study #1 which demonstrates the effectiveness of disulfiram in the treatment of alcoholism and study #2 which demonstrates the benefits of motorcycle helmet laws? Not much, I suppose. But it hardly follows that the integration of study #1 on lysergide treatment of alcoholism and study #2 on "controlled drinking" is meaningless; one is understandably concerned with which treatment has a greater cure rate. Is the essential difference between the two examples that in the former case the problems addressed by the studies are different but the problem is the same in the latter example? "Problem" is no better defined than "study" or "findings," and invoking the word clarifies little. It is easy to imagine the Secretary for Health comparing fifty studies on alcoholism treatment with fifty studies on drug addiction treatment or a hundred studies on the treatment of obesity. If the two former groups of studies are negative and the latter is positive, the Secretary may decide to fund only obesity treatment centers. From the Secretary's point of view, the problem is public health, not simply alcoholism or drug addiction treatment.

Suppose that a researcher wished to integrate existing studies on computer-assisted instruction (CAI) and cross-age tutoring (CAT) to obtain some notion of their relative effectiveness. That studies #1 and #2 on CAI used different standardized achievement tests to measure progress in mathematics is a difference that should cause little concern, considering the basic similarity of most standardized achievement tests. He who would object to integrating the findings from these two studies must face a succession of difficult questions which begin with whether he will accept as comparable two studies using different forms of the same test or whether he will accept as equal two average scores which were achieved by different patterns of item responses to the same form of the same test.

Imagine further that of 100 CAI studies, 75 were in math and 25 in science, whereas of the 100 CAT studies, 25 were in math and 75 were in science. Are the aggregated data on effectiveness from 100 studies each of CAI and CAT meaningfully comparable? It depends entirely on the exact form of the question being addressed. If CAI is naturally much more frequently applied to math instruction than to science (and vice versa for CAT), then the simple aggregation of effectiveness measures may most meaningfully answer the question of what benefits could be expected by a typical school from installing CAI (and using it in the natural manner, which means three times more extensively in math than in science) instead of instigating CAT. If, however, one were more interested in the question of whether CAI was a more effective medium than CAT, then such a comparison ought not to be confounded with problems of the difficulties of learning math versus science. In these circumstances, a straightforward aggregation of the findings in each set of 100 studies would not be most meaningful. To compare the media independently of subject taught, one could calculate effectiveness measures separately for math and science within either CAI or CAT. Then total effectiveness measures for CAI and CAT would be constructed by some appropriate method of proportional weighting.

There exists another respect in which critics are inconsistent who criticize meta-analysis as meaningless because it mixes apples and oranges. These same critics, researchers themselves, habitually perform data analyses in their own research in which they lump together (average or otherwise aggregate in analyses of variance, *t*-tests and whatever) data from different persons. These persons are as different and as much like apples and oranges in their way as studies are different from each other. Yet the same critics who object to pooling the findings of studies 1, 2, ..., 10 see nothing at all objectionable in pooling the results from persons 1, 2, ..., 100 in their own research. An inconsistancy of no small order must be acknowledged at this point, or else the critic of meta-analysis must argue convincingly that the two kinds of aggregating identified are qualitatively different; and he should specify how they are different and why it matters, which will necessarily entail presenting empirical evidence to demonstrate that studies using different populations, measuring instruments, data analyses, etc. are fundamentally incommensurable. (The ironic dilemma posed here is that such an empirical demonstration would be of itself an analysis of exactly the type which we have referred to as a "meta-analysis".)

Criticism #2. The meta-analysis approach "advocates low standards of judgment" of the quality of studies.

Although Eysenck (1978) saw us as "advocating" low standards of research quality, other critics have viewed us merely as being incapable of telling the difference between "good" and "bad" studies. We have been accused of relying on undiscriminating volume of data rather than on quality of design and evidence. In the academic wars waged over the questions of the benefits of psychotherapy, the judgment of "quality of design and evidence" has usually been the *ad hoc* impeaching on methodological of the studies of one's enemies.

Somewhere in the history of the social sciences, research criticism took an unhealthy turn. It became confused with research design. The critic often reads a published study and second guesses the aspects of measurement and analysis that should have been anticipated by the researcher. If a study "fails" on a sufficient number of these criteria--or if it fails to meet conditions of which the critic is particularly fond--the study is discounted or eliminated completely from consideration. Research design has a logic of its own, but it is not a logic appropriate to research integration. The researcher does not want to perform a study deficient in some aspect of measurement or analysis, but it hardly follows is the same in the latter example? "Problem" is no better defined than "study" or "findings," and invoking the word clarifies little. It is easy to imagine the Secretary for Health comparing fifty studies on alcoholism treatment with fifty studies on drug addiction treatment or a hundred studies on the treatment of obesity. If the two former groups of studies are negative and the latter is positive, the Secretary may decide to fund only obesity treatment centers. From the Secretary's point of view, the _problem_ is public health, not simply alcoholism _or_ drug addiction treatment.

Suppose that a researcher wished to integrate existing studies on computer-assisted instruction (CAI) and cross-age tutoring (CAT) to obtain some notion of their relative effectiveness. That studies #1 and #2 on CAI used different standardized achievement tests to measure progress in mathematics is a difference that should cause little concern, considering the basic similarity of most standardized achievement tests. He who would object to integrating the findings from these two studies must face a succession of difficult questions which begin with whether he will accept as comparable two studies using _different_ forms of the _same_ test or whether he will accept as equal two average scores which were achieved by _different_ patterns of item responses to the _same_ form of the _same_ test.

Imagine further that of 100 CAI studies, 75 were in math and 25 in science, whereas of the 100 CAT studies, 25 were in math and 75 were in science. Are the aggregated data on effectiveness from 100 studies each of CAI and CAT meaningfully comparable? It depends entirely on the exact form of the question being addressed. If CAI is naturally much more frequently applied to math instruction than to science (and vice versa that after a less-than-perfect study has been done, its findings should not be considered. A logic of research integration could lead to a description of design and analysis features and study of their covariance with research findings. If, for example, the covariance is quite small between the size of an experimental effect and whether or not subjects were volunteers, then the force of the criticism that some experiments used volunteers is clearly diminished.

Our early work on the effects of psychotherapy (Smith and Glass, 1977) never strayed far from a sensitivity to design and methods in the studies integrated. However, across the field of psychotherapy outcome evaluation, there was basically no correlation between the "quality" (in the sense of Campbell and Stanley, 1966, and others) of the design and the size of psychotherapy effect (Smith and Glass, 1977, p. 758, Table 4). Thus any distinctions between "good" and "bad" studies would leave the overall picture unchanged--a fact that should be clear to anyone who understands what the absence of correlation implies. No purpose would have been served by reporting results separately for "good" and "bad" studies since they would have been essentially the same. In a meta-analysis of educational research on the effect of class-size on achievement, Glass and Smith (1979) found that quality of research design (essentially the degree of control exercised over the assignment of pupils to classes) was the highest correlate of effects. The sensible course was elected, and results were presented only for the studies in which careful experimental control was exercised.

An early attempt at meta-analysis was characterized somewhat cynically by a critic as follows: "Although no single study was well enough done to prove that psychotherapy is effective, when you put all these bad studies together, they show beyond doubt that therapy works." This skeptical characterization with its paradoxical ring is a central thesis of research integration. In fact, many weak studies can add up to a strong conclusion. Suppose that, in a group of 100 studies, studies 1-10 are weak in representative sampling but strong in other respects; studies 11-20 are weak in measurement but otherwise strong; studies 21-30 are weak in internal validity only; studies 31-40 are weak only in data analysis; and so on. But imagine also that all 100 studies are somewhat similar in that they show a superiority of the experimental over the control group. The critic who maintains that the total collection of studies does not support strongly the conclusion of treatment efficacy is forced to invoke an explanation of multiple causality (i.e., the observed difference can be caused either by this particular measurement flaw _or_ this particular design flaw, _or_ this particular analysis flaw, _or_...). The number of multiple causes which must be invoked to counter the explanation of treatment efficacy can be embarrassingly large for even a few dozen studies. Indeed, the multiple-defects explanation will soon grow into a conspiracy theory or else collapse under its own weight. Respect for parsimony and good sense demands an acceptance of the notion that imperfect studies can converge on a true conclusion.

An important part of every meta-analysis with which we have been associated has been the recording of methodological weaknesses in the original studies and the examination of their covariance of study findings. Thus, the influence of "study quality" on findings has been regarded consistently as an empirical _a posteriori_ question, not an _a priori_ matter of opinion or judgment used in excluding large numbers of studies from consideration. But a critic once asked us, "Why do you study the difference in the findings of 'good' vs. 'bad' studies? If you found a difference, wouldn't you reject the 'bad' studies? And if you found no difference, wouldn't the findings of the 'good' studies be the same as those for all studies regardless of quality?" The dilemma was neatly posed, and we hope the answer is comprehensible. Surely, the "good" studies (i.e., those with excellent controls and sophisticated technology) are to be believed if a conflict is observed between findings of good and poor studies (cf. Glass and Smith, 1979). However, if "good" and "poor" studies do not differ greatly in their findings, a large data base (all studies regardless of quality) is much to be preferred over a small data base (only the "good" studies). The larger data base can be more readily subdivided to answer specific sub-questions that are inevitably provoked by the answers to the general questions (e.g., "But are behavioral therapies superior to cognitive therapies for children with low I.Q.?"). The smaller data base of "good" studies only is likely to have too few instances to address many sub-questions. Moreover, even when the results of "good" and "bad" studies differ, even the bad or not-so-bad studies can be informative; for suppose that six studies of quality "10" on a ten-point scale show a correlation of X and Y of .70 on the average, and that twelve studies of quality "9" show an _r_ of .65, studies of quality "8" an _r_ of .60, and so on down to quality "1" an _r_ of .10, say. This pattern is far more informative and lends greater credence to a _r_ of .70 for six studies of top qaulity than would the results of the six studies in isolation from all others.

The covariation of research quality with results is, then, an empirical matter of central concern in meta-analysis, as well as being of interest to research methodologists who find meta-analysis too much to swallow. Fortunately, we have several thousand data that can inform us on the general question.

In Table 7.1 appears a summary of the differences in results among studies of varying research quality for twelve different meta-analyses. Each meta-analysis was performed on a literature of comparative experimental findings. The basic unit of measurement for the meta-analysis was the effect size, ES, and in each instance it was defined so that positive values indicated findings in accord with the favored hypothesis of the field in question (e.g., a positive ES in Hartley's meta-analysis of computer assisted math instruction indicated a superiority of CAI over traditional teaching). In each meta-analysis, the rating of High, Medium, or Low research quality was primarily an assessment of internal validity of the experiment (Campbell and Stanley, 1966).

If Table 7.1 achieves nothing else, it ought to be, at the very least, an effective antidote to rampant _a priorism_ on the matter of which studies should be admitted as evidence in deciding research questions. Some of the meta-analyses in Table 7.1 show a relationship between design quality and findings and others do not. But in those analyses with substantial numbers of cases, the differences in size of average experimental effects between High validity and Low validity experiments are surprisingly small. The only notable exceptions to this trend in the entire table are Hartley's ('78) tutoring analysis, Smith ('80) and Carlberg's ('79) resource room analysis; but in each of these instances, as just suggested, the large deviations are probably merely the consequence of small _n_'s in particular categories. As a general rule, there is seldom much more than one-tenth standard deviation difference between average effects for High validity and Low validity experiments.

Table 7.1

Relationship Between Research Quality (Internal Validity) and Findings in 12 Meta-Analyses of Experimental Literatures

		Relationship Between Internal Validity and Average Experimental Effect Size		
Investigator(s)	Topic	High	Medium	Low
Hartley ('77)	Computer-based Instruction	n: 11	55	23
		$\bar{\Lambda}$:: .311	.389	.503
	Tutoring	n: 52	12	9
		$\bar{\Delta}$.: .584	.306	1.066
Kulik, Kulik & Cohen ('79)	Individual Instruction	n: 22		22
		$\bar{\Delta}$.: .409		.604
Smith ('80a)	Sex bias in psychotherapy	n: 30	26	4
		$\bar{\Delta}$.: -.18	-.01	.77
Smith ('80b)	Effects of aesthetic educ. on basic skills	n: 84	48	117
		$\bar{\Delta}$.: .53	.52	.59
Carlberg ('79)	Spec. ed. room placement vs. reg. room placement	n: 83	187	52
		$\bar{\Delta}$.: -.19	-.11	.02
	Resource room placement vs. reg. room placement	n: 3	31	5
		$\bar{\Delta}$.: 1.13	.12	.56
	Spec. educ. intervention vs. classroom treatment	n: 40	81	35
		$\bar{\Delta}$.: .19	.27	.53

Table 7.1 (continued)

Miller ('78)	Drug therapy for psych. disorders	n:	297	16	37
		$\bar{\Delta}$.:	.48	.54	.64
Hearold ('79)	Effects of TV on anti-social behav.	n:	176	176	176
		$\bar{\Delta}$.:	.33	.30	.27
	Effects of TV on "pro-social" behav.	n:	35	35	35
		$\bar{\Delta}$.:	.59	.63	.67
			High	Medium	Low
SUBTOTALS		n:	833	667	515
		$\bar{\Delta}$.:	.36	.21	.43
Smith, Glass & Miller ('80)	Psychotherapy	n:	1157	378	224
		$\bar{\Delta}$:	.82	.75	.68
TOTALS		n:	1990	1045	739
		$\bar{\Delta}$:	.63	.40	.51

Our experience with meta-analyses of experiments was matched by Yin, Bingham and Heald (1976) in their study of the relationship between case study quality and findings. Yin and his colleagues collected 140 case studies on technological innovations, every study they could find that appeared after 1965. They devised four criteria for judging the quality of the studies: 1) presence of operational measures of innovative device and outcomes, 2) presence of some relevant research design, 3) overall adequacy of evidence in relation to conclusions, and 4) adequacy of evidence in relation to each stated outcome. They correlated research quality, so defined, with study outcomes and concluded:

> "To the extent that one objective of our investigation was to examine the widest possible range of reported innovative experiences, there was thus strong reason not to discard the lower quality studies. At the same time, the general lack of relationship between quality and the outcomes of the innovative experience suggested that the inclusion of lower quality studies would not affect the overall conclusions to be drawn from the review." (Yin, Bingham and Heald, 1976, pp. 153-4)

In an earlier study (Yin and Yates, 1975), the investigators did observe an association between research quality and findings, just as we see a relationship in some literatures and not in others. Without thinking about the matter further, one is tempted to ask why "poor quality" studies are included in the first place if they'll only be retained provided they agree in their findings with the high quality studies. If there were virtually huge numbers of both well-done and poorly-done studies on a question, the answer would be clear: throw away the poorly-done studies and heed the message of the high quality research. But the usual situation is that there exist several studies, some of which are high quality, some average and some poor.

Suppose that of fifty experiments on the effects of jogging on life expectancy, 25 are judged to be of poor design and execution, 15 are regarded as moderately well done and 10 are well-done. Suppose further that the average effect (experimental vs. control group difference) is 2.86 years life expectancy favoring the experimental group in the 10 best designed studies. Should one base his opinion on the results of these 10 studies and ignore the findings of the other forty? Let's press on and see. Suppose that the effects shown by the 15 moderately well done and 25 poorly done experiments were 2.74 years and 2.60 years, respectively. These findings do, in fact, support the finding of the less numerous well done studies and make it more credible. Imagine contrariwise that the average effects for the moderately well done and poorly done experiments were -0.47 years and 8.65 years, respectively. Now the finding of the ten well done experiments is placed in a context of chaotic error and variability and it is more suspect. People reason and judge with the help of complex patterns and contexts; scholars who are doctrinaire about research quality when they integrate research studies ignore this fact. It is precisely this fact that was ignored in a widely publicized critique of our meta-analysis of the school class-size and achievement relationship (Educational Research Service, 1980).

Criticism #3. Selection Bias in Reported Research.

Meta-analysis is dependent on the findings that researchers report. Its findings will be biased if, as is surely true, there are systematic differences among the results of research that appear in journals vs. books vs. theses vs. unpublished papers.

The findings of a dozen meta-analyses can be used to inform us on the severity of one aspect of this criticism. Several investigators working on the integration of experimental literatures compared the effects revealed by experiments depending of whether they were published in journals, books, doctoral or master's theses, or not published at all. The results are tabulated as Table 7.2.

The findings in Table 7.2 are fairly consistent. In every one of the ten instances in which the comparison can be made, the average experimental effect from studies published in journals is larger than the corresponding effect estimated from theses and dissertations. That is, if one integrates only "published" (meaning journal published) studies, the impression of support for the favored hypothesis is artificially enhanced over what would be seen if the entire literature were integrated (i.e., journals, books and dissertations). The bias in the journal literature relative to the bias in the dissertation literature is not inconsiderable. The mean effect size for journals is .64 as compared with .48 for the dissertation literature; hence, the bias is of the order of $[(.64 - .48)/.48]\ 100\% = 33\%$. Thus, findings reported in journals are, on the average, one-third standard deviation more favorably disposed toward the favored hypotheses of the investigators than findings reported in theses or dissertations.

Comparisons of average effect sizes among other sources of publication

Table 7.2

Relationship Between Source of Publication and Findings in 12 Meta-Analyses of Experimental Literatures

			Source of Publication			
Investigator(s)	Topic		Journal	Book	Thesis	Unpubl.
Kavale ('79)	Psycholinguistic training	n:	13		16	5
		$\bar{\Delta}$.:	.50		.30	.37
Hartley ('77)	Computer-based instruc.	n:	34		13	34
		$\bar{\Delta}$.:	.36		.28	.54
	Tutoring	n:	9		47	17
		$\bar{\Delta}$.:	.77		.40	1.05
Rosenthal ('76)	Experimenter bias	n:	25		50	
		$\bar{\Delta}$.:	1.02		.74	
Smith ('80a)	Sex bias in psychotherapy	n:	28		32	
		$\bar{\Delta}$.:	.22		-.24	
Smith ('80b)	Effects of aesthetics educ. on basic skills	n:	29		164	56
		$\bar{\Delta}$.:	1.08		.48	.50
Carlberg ('79)	Spec. ed. room placement vs. reg. room placement	n:	146	17	45	114
		$\bar{\Delta}$:	-.09	-.01	-.16	-.14
	Resource room plac. vs. reg. room place.	n:	33	6		
		$\bar{\Delta}$.:	.32	-.09		

Table 7.2 (continued)

Miller ('79)	Drug therapy of psych. disorders	n:	336	21		
		$\bar{\Delta}$.:	.49	.56		
Hearold ('79)	Effects of TV on anti-social behav.	n:	262	120	96	13
		$\bar{\Delta}$:	.40	.14	.18	.23
SUBTOTALS		n:	1025	177	473	268
		$\bar{\Delta}$:	.38	.18	.30	.27
Smith, Glass & Miller ('80)	Psychotherapy	n:	1179	42	483	61
		$\bar{\Delta}$.:	.87	.80	.66	1.96
TOTALS		n:	2204	219	956	329
		$\bar{\Delta}$.:	.64	.30	.48	.58

are less clear, in part perhaps, because of the ambiguity in labels such as "Unpublished" or "book." In four of six instances, journals gave more favorable results than books. In four of eight instances, the average effect size for journals was larger than for unpublished studies. Unpublished studies seemed to divide along the following lines: one large group of old unpublished studies containing unremarkable results that never caught anyone's attention, and a smaller group of new studies circulating through the "invisible college" while waiting to be published.

White (1976) also produced evidence of a selective publication effect in his meta-analysis of the relationship between socio-economic status and achievement. The average of 165 correlations published in books was .31; 38 $\underline{r}$'s in journals averaged .25, and 286 dissertation correlations between achievement and SES showed an average of .20. This trend, toward weaker relationships in dissertations than in journals, agrees with the trend established above for various experimental literatures.

The compilation of results from various meta-analyses shows that there is substance to the criticism that most disciplines show evidences of a selection bias in what they publish. And the bias may be large in some instances. Smith's (1980) meta-analysis of sex-bias in psychotherapy is particularly relevant, as a final example. The very direction of the bias was reversed between the dissertation literature and published journals (from demonstrating a bias in favor of women in the thesis literature to a bias against women in journals); that this reversal was in accord with political ideologies that are presumed to control access to journals makes the case even stronger that disciplines are prone to the temptation to reward findings they approve of by publishing them in more prestigious places.

However, the fact of the existence of selective publication tendencies is not in itself a cogent criticism of meta-analysis, which after all, is used here to demonstrate the existence and the magnitude of the phenomenon. Indeed, the problem of selective publication cannot be dealt with adequately in integrating a research literature except by meta-analytic means, i.e., by collecting all of the literature at the outset and analyzing it separately by mode of publication.

There exists another factor with respect to which selection often takes place during research integration, namely, the date on which the studies were published. It is common for reviewers to restrict their attention to a particular span of years and review only studies of that period, e.g., "This review will consider all laboratory studies on attention processes published after 1960." The choice of dates is invariably arbitrary and governed by convenience. It behooves us to inquire into the matter of chronological trends in research findings.

In Table 7.3 appears a compilation of correlations between date of publication and effect size from size meta-analyses of experimental literatures.

The average of the eight correlations in Table 7.3 is +.13, indicating that more recently published experiments show a slight tendency toward larger effects than older studies. (The weighted average r, each r weighted by the number of effect sizes in the particular meta-analysis, equals +.07. The unweighted average is probably more sensible because it is not affected by some meta-analyses arbitrarily having more data points.) Assuming a correlation of +.13 between date of publication and effect size and some reasonable parameters for the independent variable (Date) and the dependent

Table 7.3

Correlation Between Date of Publication and Effect Size for Six Meta-Analyses of Experimental Literatures

Investigator(s)	Topic	Correlation Between Date of Publication and Effect Size
Kavale ('79)	Psycholinguistic training	r = -.01 (n = 25)
Hall ('78)	Gender effects in non-verbal coding	r = .28 (n = 44)
Smith ('80)	Sex bias in psychotherapy	r = .29 (n = 60)
Carlberg ('79)	Spec. ed. room placement	r = .02 (n = 322)
	Resource room placement	r = .32 (n = 39)
	Other spec. ed. intervention	r = .08 (n = 156)
Miller ('78)	Drug treatment of psychological disorders	r = -.01 (n = 358)
Smith, Glass & Miller ('80)	Psychotherapy	r = .07 (n = 1,764)

variable (Effect Size or Δ), then a linear regression equation can be constructed that relates date of publication to effect size:

$$\hat{\Delta} = .13\left(\frac{.67}{4}\right)\text{Date} + .70 - .13\left(\frac{.67}{4}\right)1970$$

The above equation contains some assumed values for the means and standard deviations of Date and Δ:

Variable	Mean	Standard Deviation
Date	1970	4 years
Δ	.70	.67

Substituting the dates 1965 and 1975, each about one standard deviation away from the mean, into the regression equation gives:

$$\hat{\Delta}_{1965} = .59, \quad \text{and}$$

$$\hat{\Delta}_{1975} = .81.$$

These calculations indicate that the typical correlation between date of publication and effect size ($\underline{r} = .13$) implies that experiments published in 1975 show a .22 average effect size advantage over experiments published in 1965. This difference amounting to $[(.81 - .59)/.59]\,100\% = 37\%$ is comparable to the difference in average effect size between journals and theses. Thus the concerns about bias that applied in the case of selectivity in publication outlet appear to apply with nearly equal force to the case of selection of studies by date. It would seem ill-advised to begin the integration of an empirical research literature by arbitrarily restricting the studies considered to those published in refereed journals after 1960, for example.

Criticism #4. Lumpy (Non-Independent) Data.

Meta-analyses are conducted on large data sets in which multiple results are derived from the same study; this renders the data non-independent and gives one a mistaken impression of the reliability of the results.

Of all the technical criticisms of meta-analysis that have been published in the last five years (and most of these criticisms are quite off-the-mark and shallow), the reminder that meta-analyses are typically carried out on lumpty sets of non-independent data is quite cogent. The principal implication of this non-independence is a reduction in the reliability of estimation of averages or of regression equations. For example, if Study #1 gave effects .2, .2, .2 and .2 and Study #2 gave effects .6, .6, and .6, one would have little reason to believe that he had been informed seven times about the aggregate result in question; rather the true "degrees of freedom" would seem to be somewhat closer to 2, the number of studies, than to 7, the number of effects. A facile solution to this problem of non-independence would be to average all findings within a study up to the level of the study and proceed with a meta-analysis with "studies" as the unit of analysis. No doubt there will be instances in which this resolution of the problem will be satisfactory. But in most instances, it is likely to obscure many important questions that can only be addressed at the "within study" level of outcome variables, say. The effect on accuracy of estimation of complex interdependencies in a meta-analysis data base was addressed at the end of Chapter Six.

CONCLUSION

Of course, it is unclear what meta-analysis will contribute to the progress of empirical research. One can imagine a future for research in the social and behavioral sciences in which questions are so sharply put and techniques so well standardized that studies would hardly need to be integrated by merit of their consistent findings. But that future seems unlikely. Research will probably continue to be an unorganized, decentralized, non-standardized activity pursued simultaneously in dozens of places without thought to how it will all fit together in the end. The need for formal techniques of research integration like those we have illustrated will probably grow. Whether future techniques will resemble these is uncertain, but we suspect they will. The approach we call meta-analysis seems to be too plainly reasonable to be false in any simple sense. Whether it will be useful is a different matter.

BIBLIOGRAPHY

Andrews, G. *A meta-analysis of the treatment of stuttering*. Paper presented at the annual meeting of the American Speech and Hearing Association, 1979.

Andrews, G., Guitar, B. and Howie, P. A meta-analysis of stuttering therapy outcome studies. *Journal of Speech and Hearing Disorders*, (in press).

Athappilly, K. K. *A meta analysis of the effects of modern mathematics in comparison with traditional mathematics in the American educational system*. Ann Arbor, MI: University Microfilms International, 1980.

Ashton, W. D. *The logit transformation*. London: Charles Griffin, 1972.

Bandura, A. On paradigms and recycled ideologies. *Cognitive Therapy and Research*, 1978, *2*, 79-103.

Barton, M. A. and Glass, G. V. *Integrating studies that have quantitative independent variables*. Paper delivered at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Barton, M. A. Two new uses of indicator variables in Meta-analysis, *Evaluation in Education: An International Review Series*, 1980, *4*, 28-30. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley, *Science*, 1975, 187, 398, 404.

Blanchard, E. B., Andrasik, F. Ahles, T.A., Teders, S. J. and O'Keefe, D. Migraine and tension headache: a meta-analytic review. Paper published by the Headache Project, 129 Milne Hall, State University of New York at Albany. 1980.

Blyth, C. R., On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 1972, 67, 364-366.

Bredderman, T. *Elementary school process curricula: A meta-analysis*. (ERIC Document Reproduction Service No. ED 170 333).

Bredderman, T. Process curricula in elementary school science. *Evaluation in Education: An International Review Series*, 1980, *4*, 43-44. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Cahen, L. S. Meta-analysis - a technique with promise and problems. *Evaluation in Education: An International Review Series*, 1980, *4*, 37-39. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Cahen, L. and Filby, N. The class size/achievement issue: New evidence and a research plan. *Phi Delta Kappan*, 1979, *60*, 492-296.

Carlberg, C. *Meta-analysis of special education treatment techniques*. Doctoral dissertation, Boulder, Colorado: University of Colorado, 1979.

Cook, T. D. The potential and limitations of secondary evaluations. Chapter 6, pp. 155-234 in Apple, M. W., Subkoviak, H. S. and Lufler, J. R. (Eds.), *Educational evaluation: Analysis and responsibility*. Berkeley: McCutchan, 1974.

Cook, T. D. and Leviton, L. C. Reviewing the literature: a comparison of traditional methods with meta-analysis. *Journal of Personality*, (in press).

Cooper, M. and Rosenthal, R. A comparison of statistical and traditional procedures for summarizing research. Evaluation in Education: An International Review Series, 1980, *4*, 33-36. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Cooper, M., Burger, M., and Good, T. L. Gender differences in learning control beliefs of young children. *Evaluation in Education: An International Review Series*, 1980, *4*, 73-75. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Cooper, H. M. and Rosenthal, R. Statistical vs. traditional procedures for summarizing research findings. *Psychological Bulletin*, 1980, *87*

Cronbach, L. J. and Furby, L. How should we measure "change" -- or should we? *Psychological Bulletin*, 1970, *74*, 68-80.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. *The Dependability of Behavioral Measurements*. NY: John Wiley, 1971.

Davidson, T. B. *Meta-analysis of the neuropsychological assessment of children*. Unpublished dissertation, University of Denver, 1978.

Edgington, E. S. An additive method for combining probability values from independent experiments. *Journal of Psychology*, 1972a, 80, 351-363.

Edgington, E. S. A normal curve method for combining probability values from independent experiments. *Journal of Psychology*, 1972b, 82, 85-89.

Educational Research Service, Class size research: a critique of recent meta-analyses. Arlington, Virginia: Educational Research Serivce, Inc., 1980. 81 pp.

El-Nemr, M. A. Meta-analysis of the outcomes of teaching biology as inquiry. Doctoral dissertation, Boulder, Colorado: University of Colorado, 1979.

Eyenck, H. J. An exercise in mega-silliness. American Psychologist, 1978, 33, 517.

Eysenck, H. J. Correspondence. Bulletin of the British Psychological Society, 1978b, 31, 56.

Feldman, K. A. Using the work of others: some observations on reviewing, integrating, and consolidating findings. In R. B. Smith, B. Anderson, and P. Manning (Eds.), Handbook of Social Science Research Methods. New York: Irvington Publishers, Inc., 1978.

Ferguson, P. C. An Integrative meta-analysis of psychological studies investigating the treatment outcomes of meditation techniques. Doctoral dissertation, University of Colorado, 1980.

Finney, D. J. Probit analysis (3rd ed.). Cambridge: Cambridge University Press, 1971.

Gage, N. L. The Scientific Basis of the Art of Teaching. NY: Teachers College Press, 1978.

Gallant, D. M. Evaluation of compulsory treatment of the alcoholic municipal court offender. In Millie, N. and Mendelson, J. (eds.), Recent Advances in Studies of Alcoholism. U.S. Government Printing Office, 1971.

Gallo, P. S. Meta-analysis-- a mixed meta-phor. American Psychologist, 1978, 33, 515-517.

Gillen, A. L. Critique of meta-analysis of research on the relationship of class-size and achievement. Unpublished paper. Ottawa, Ontario: Gillen Associates, May, 1979.

Gilligan, J. Review of literature. In Breenblatt, M.; Solomon, M.; Evans, A.; and Brooks, G. (eds.), Drug and Social Therapy in Chronic Schizophrenia. Sprinfield, ILL.: C.C. Thomas, 1965.

Glass, G. V. Integrating findings: the meta-analysis of research. Review of Research in Education, 1978, 5, 351-379.

Glass, G. V. Primary, secondary and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G. V. Reply to Mansfield and Bussey. Educational Researcher, 1978, 7, 3.

Glass, G. V., et al. Teacher "indirectness" and pupil achievement: An integration of findings. Boulder: University of Colorado, Laboratory of Educational Research, 1977.

Glass, G. V., & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. American Educational Research Journal, 1969, 6. 403-414.

Glass, G. V. and Smith, M. L. Meta-analysis of research on the relationship of class-size and achievement. Evaluation and Policy Analysis, 1979, 1, 2-16.

Glass, G. V. and Smith, M. L. Reply to Eysenck. American Psychologist, 1978, 33, 517-518.

Glass, G. V. & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Group for the Advancement of psychiatry. Pharmacotherapy and Psychotherapy: Paradoxes, Problems and Progess 9 (March 1975).

Haertel, G. D., Haertel, H. Classroom socio-psychological environment. Evaluation in Education: An International Review Series, 1980, 4, 113-114. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Haertel, G. D., Walberg, H. J. and Haertel, E. H. Social-psychological environments and learning: A quantitative synthesis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Hall, J. A. Gender effects in decoding nonverbal cues. Psychological Bulletin, 1978, 85, 845-857.

Hartley, S. S. Instruction in mathematics. Evaluation in Education: An International Review Series, 1980, 4, 56-57. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Hartley, S. S. Meta-analysis of the effects of individually paced instruction in mathematics. Doctoral dissertation, Boulder, Colo.: University of Colorado, 1977.

Hastings, N. A. J. and Peacock, J. B. Statistical Distributions. London: Butterworth, 1974.

Hearold, S. Meta-analysis of the effects of television on social behavior. Doctoral dissertation, Boulder, Colo.: University of Colorado, 1979.

Hearold, S. L. Television and social behavior. Evaluation in Education: An International Review Series, 1980, 4, 94-95. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Hedges, L. V. Combining the results of experiments using different scales of measurement. Doctoral dissertation, Stanford University, 1980.

Hedges, L. V. Unbiased estimation of effect size. Evaluation in Education: An International Review Series, 1980, 4, 25-27. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Hedges, L. V. and Olkin, I. Three vote-counting methods for the estimation of effect size and statistical significance of combined results. Psychological Bulletin (in press).

Hekmat, H. Systematic versus semantic desensitization and implosive therapy: a comparative study. Journal of Consulting and Clinical Psychology, 1973, 40, 202-209.

Hess, F. Class size revisited: Glass and Smith in perspective. Syracuse, New York: Minoa Central Schools, 1979 (ERIC Document Reproduction Service No. ED 168 129).

Horan, P. F. and Lynn, D. D. Learning hierarchies research. Evaluation in Education: An International Review Series, 1980, 4, 82-83. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Hunter, J. E. Cumulating results across studies: correction for sampling error, a proposed moratorium on the significance test, and a critique of current multivariate reporting practice. East Lansing, Mich.: Dept. of Psychology, Michigan State University, 1979. (Unpublished) 28 pp.

Hunter, J. E. and Schmidt, F. L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Psychology, 1978, 63, 1-11.

Hunter, J., Schmidt, F., and Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Iverson, B. K. and Walberg, H. J. Home environment. Evaluation in Education: An International Review Series, 1980, 4, 107-108. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Iverson, B. K. and Walberg, H. J. Home environment and learning: A quantitative synthesis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Jackson, G. B. Methods for reviewing and integrating research in the social sciences. Final Report to the National Science Foundation for Grant No. DIS 76-20309. Washington, D.C.: Social Research Group, George Washington University, April, 1978.

Jacobs, J. A. and Critelli, J. W. Treatment-outcome interactions. Evaluation in Education: An International Review Series, 1980, 4, 31-32. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Johnston, J. Econometric Methods (2nd edition). N.Y.: McGraw-Hill, 1972.

Jones, L. V., & Fiske, D. W. Models for testing the significance of combined results. Psychological Bulletin, 1953, 50, 375-382.

Kavale, K. The effectiveness of psycholinguistic training: a meta-analysis. Riverside, Calif.: University of California-Riverside, 1979.

Kavale, K. Meta-analysis of experiments on the treatment of hyperactivity in children. Riverside, Calif.: University of California-Riverside, 1980.

Kavale, K. Psycholinguistic training. Evaluation in Education: An International Review Series, 1980, 4, 88-89. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Kavale, K. and Carlberg, C. Regular versus special class placement for exceptional children. Evaluation in Education: An International Review Series, 1980, 4, 91-93. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Kavale, K. The relationship between auditory perceptual skills and reading ability: a meta-analysis. Riverside, Cal.: School of Education, University of California-Riverside, 1980.

Kazdin, A. E. and Rogers, T. On paradigms and recycled ideologies: Analogue research revisited. Cognitive Therapy and Research, 1978, 2, 105-117.

Kerlinger, F. N. Foundations of Behavioral Reserach. NY: Holt, Rinehart & Winston, 1964.

Kish, L. Survey Sampling. NY: Wiley, 1965.

Kraemer, H. C. and Andrews, G. A non-parametric technique for meta-analysis effect size calculation. Stanford University. Unpublished paper, 1980.

Krol, R. A. A meta analysis of comparative research on the effects of desegregation on academic achievement. Ann Arbot, MI.: University Microfilms International, 1979.

Krumboltz, J. D. and Thoresen, C. E. The effect of behavioral counseling in group and individual settings on information-seeking behavior. Journal of Counseling Psychology, 1964, 11, 324-333.

Kulik, J. A., Kulik, C. C. amd Cohen, P. A. A meta-analysis of outcome studies of Keller's personalized system of instruction. American Psychologist, 1979, 34, 307-318.

Kulik, J. A. and C. Kulik. Individualised college teaching. Evaluation in Education: An International Review Series, 1980, 4, 64 67. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Light, R. J. Capitalizing on variation: How conflicting research findings can be helpful for policy. Educational Researcher, 1979, 8, No. 9.

Light, R. J. Synthesis methods: some judgement calls that must be made. Evaluation in Education: An International Review Series, 1980, 4, 13-17. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Light, R. J. and Smith P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Luborsky, L., Singer, B., and Luborsky, L. Comparative studies of psychotherapies. Archives of General Psychiatry 32 (1975): 995-1008.

Luiten, J. W. Advance organisers in learning. Evaluation in Education: An International Review Series, 1980, 4, 49-50. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Luiten J., Ames, W. and Ackerson, G. The advance organizer: A review of research using Glass's technique of meta-analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Lynn, D. D., and Donovan, J. M. Medical versus surgical treatment of coronary artery disease. Evaluation in Education: An International Review Series, 1980, 4, 98-99. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Lysakowski R. S., and Walberg H. J. Classroom reinforcement. *Evaluation in Education: An International Review Series*, 1980, *4*, 115-116. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford Cal.: Stanford University Press, 1974.

Mansfield, R. S. and Bussey, T. V. Meta-analysis of research: A rejoinder to Glass. *Educational Researcher*, 1977, *6*, 3.

May, P. R. A. Psychotherapy and ataraxic drugs. In Bergin, A. E. and Garfield, S. L. (eds.), *Handbook of Psychotherapy and Behavior Change*, New York: Wiley, 1971.

McCance, C., and McCance, P. F. Alcoholism in north-east Scotland: Its treatment and outcome. *British Journal of Psychiatry* 115 (1969): 189-98.

McGaw, B. and Glass, G. V. *Choice of the metric for effect size in meta-analysis*. *American Educational Research Journal*, In Press.

Miller, T. I. Drug therapy for psychological disorders. *Evaluation in Education: An International Review Series*, 1980, *4*, 96-97. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Miller, T. I. *The effects of drug therapy on psychological disorders*. Doctoral Dissertation, Boulder, Colo.: University of Colorado, 1977.

Mood, A. M. On the asymptotic power efficiency of certain non-parametric two-sample tests. *Annals of Mathematical Statistics*, 1954, *25*, 514-522.

Mosteller, F. M., and Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of social psychology: Volume I. Theory and method. Cambridge, Mass.: Addison-Wesley, 1954.

Mosteller, F. M., and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronson (Eds.), Handbook of social psychology (2nd ed.) Reading, Mass.: Addison-Wesley, 1968.

Pacht, A. R., Bent, R., Cook, T. D., Klebanoff, L. B., Rodgers, D. A., Sechrest, L., Strupp, H., and Theaman, M. Data-based meta analyses as a tool in literature reviews. Journal of Personality. In press.

Paul, G. L. and Licht, M. H. Resurrection of uniformity assumption myths and the fallacy of statistical absolutes in psychotherapy research. Journal of Consulting and Clinical Psychology, 1978, 46, 1531-1534.

Pearlman, K. The validity of tests used to select clerical personnel: A comprehensive summary and evaluation (Technical Study TS-79-1). Washington D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center, August 1979. (NTIS No. PB 80-102650).

Peterson, P. L. Direct and open instructional approaches: effective for what and for whom? Working Paper No. 243 of the Wisconsin Research and Development Center for Individualized Schooling, University of Wisconsin, Madison, October, 1978.

Peterson, P. L. Open versus traditional classrooms. Evaluation in Education: An International Review Series, 1980, 4, 58-60. (Available from Pergamon Press, Maxwell House/Fairview Park, Elmsford, NY 10523.)

Posavac, E. J. Evaluations of patient education programs: a meta-analysis Evaluation and the Health Professions, 1980, 3, 47-62.

Presby, S. Overly broad categories obscure important differences between therapies. American Psychologist, 1978, 33, 514-515.

Price, G. E., and Borgers, S. B. An evaluation of the sex-stereotyping effect as related to counselor perceptions of courses appropriate for high school students. Journal of Counseling Psychology, 1977, 24, 240-243.

Redfield, D. L. and Rousseau, E. W. Meta-analysis of experimental research findings on teacher questioning behavior. Unpublished manuscript, University of Arizona, 1979.

Roid, G., Brodsky, G. and Bigelow, D. A. Meta-analysis in mental health evaluation research. Unpublished paper from the Oregon Mental Health Division; Salem, Oregon, 1979. 34 pp.

Rosenthal, R. Combining probabilities and the file drawer problem. Evaluation in Education: An International Review Series, 1980, 4, 18-21. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Rosenthal, R. Combining results of independent studies. Psychological Bulletin, 1978, 85, 185-193.

Rosentahl, R. Experimenter effects in behavioral research. New York: Irvington Publishers, 1976.

Rosenthal, R. The "file drawer problem: and tolerance for null results. Psychological Bulletin, 1979, 86, 638-641.

Rousseau E. W., and Redfield, D. L. Teacher questioning. Evaluation in Education: An International Review Series, 1980, 4, 51-52. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Schlesinger, H. J., Mumford, E. and Glass, G. V. A critical review and indexed bibliography of the literature up to 1978 on the effects of psychotherapy on medical utilization. Denver, Colo.: Department of Psychiatry, University of Colorado Medical Center, 1978.

Schlesinger, H. J., Mumford, E. and Glass, G. V. Effects of psychological intervention on recovery from surgery. Chapter 2 in Guerra, F. and Aldrete, J. A. (Eds.), Emotional and Psychological Responses to Anesthesia and Surgery. NY: Grune & Stratton, 1980.

Schwab, D., Olian-Gottlieb, J. and Heneman III, H. Between-subjects expectancy theory research: A statistical review of studies predicting effort and performance. Psychological Bulletin, 1979, 86(1). 139-147.

Shah, I. Caravan of dreams. Baltimore: Penquin Books, 1968.

Shaprio, D. A. and Shapiro, D. The "double standard" in evaluation of psychotherapies. Bulletin of the British Psychological Society, 1977, 30, 209-210.

Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. I and II. Psychometrika, 1962, 27, 125-140 and 219-246.

Simpson, E. H. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B., 1951, 13, 238-41.

Simpson, S.N. Comment on "Meta-analysis of Research on Class Size and Achievement". Educational Evaluation and Policy Analysis,1980, 2, 81-83.

Smith, M. L. Effects of aesthetics education on basic skills learning. Unpublished report. Boulder, Colo.: Laboratory of Educational Research, University of Colorado, 1980a.

Smith, M. L. Publication bias and meta-analysis. Evaluation in Education: An International Review Series, 1980, 4, 22-24. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Smith, M. L. Sex bias in counseling and psychotherapy. Psychological Bulletin, 1980b, 87, 392-407.

Smith, M. L. Teacher expectations. Evaluation in Education: An International Review Series, 1980, 4, 53-55. (Available from Pergamon Press, Maxwell House,Fairview Park; Elmsford, NY 10523.)

Smith, M. L. and Glass, G. V. Class-size and its relationship to attitudes and instruction. Boulder, Colo.: Laboratory of Educational Research, University of Colorado, July, 1979.

Smith, M. L. and Glass, G. V. Meta-analysis of psychotherapy outcome studies. American Psychologist, 1977, 32, 752-760.

Smith, M. L. and Glass, G. V. and Miller, T. I. Benefits of psychotherapy. Baltimore, Md.: John Hopkins University Press, 1980.

Snedecor, G. W., and Cochran, W. G. Statistical methods. (6th ed). Ames, Iowa State University Press, 1967.

Taveggia, T. C. Resolving research controversy through empirical cumulation: Toward reliable sociological knowledge. Sociological Methods & Research, 1974, 2, 395-407.

Torgerson, W. S. Theory and Methods of Scaling. NY: John Wiley, 1958.

Torrance, E. P. amd Parent, E. Characteristics of mathematics teachers that affect students' learning. Cooperative Research Project No. 1020, U.S. Office of Education. 1966.

Tukey, J. W. Exploratory Data Analysis. Reading, Mass.: Addison-Wesley, 1977.

Uguroglu, M. E. and Walberg, H. J. Motivation. Evaluation in Education: An International Review Series, 1980, 4, 105-106. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Uguroglu, M. E. and Walberg, H. J. Motivation and achievement: a quantitative synthesis. Paper presented at the annual meeting of the American Psychological Association, Toronto, 1978.

Uhlenhuth, E. H., Lipman, R. S. and Covi, L. Combined pharmacotherapy and psychotherapy. Journal of Nervous and Mental Diseases 148 (1969): 52-64.

Underwood, B. J. Interference and forgetting. Psychological Review, 1957, 64, 49-60.

Walberg, H. J. Reflections on research synthesis. Unpublished manuscript. Chicago: University of Illinois-Chicago Circle, 1978.

Walberg, H. J. and Haertel, E. H. Research integration: An introduction and overview. Introduction to special issue of Evaluation in Education: International Progress, In press.

Walberg, H. J. and Haertel, E. H. Research integration: introduction and overview. Evaluation in Education: An International Review Series, 1980, 4, 5-10. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

White, K. The relationship between socioeconomic status and academic achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

White, K. R. Socio-economic status and academic achievement. Evaluation in Education: An International Review Series, 1980, 4, 79-81. (Available from Pergamon Press, Maxwell House/Fairview Park; Elmsford, NY 10523.)

Winne, P. H. Experiments relating teacher's use of higher cognitive questions to student achievement. Review of Educational Research, 1979, 49, 13-50.

Willson, V. L. Critical values of the rank-biserial correlation coefficient. Educational and Psychological Measurement, 1976, 36, 297-300.

Willson, V. L. A meta-analysis of the relationship between student attitude and achievement in secondary school science. College of Education, Texas A & M University, College Station, Texas, 1980.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Wolins, L. Responsibility for raw data. *American Psychologist*, 1962, *17*, 657-658.

Yeany, R. H. and Miller, P. A. The effects of diagnostic/remedial instruction on science learning: a meta-analysis. Athens, GA: Department of Science Education, University of Georgia. April, 1980.

Yin, R. K. and Yates, D. Street-level governments: assessing decentralization and urban services. Lexington, Mass.: D. C. Heath, 1975.

Yin, R. K., Bingham, E. and Heald, K. A. The difference that quality makes: the case of literature reviews. *Sociological Methods and Research*, 1976, *5*, 139-156.

APPENDIX A

Coding Form Used in the Psychotherapy Meta-analysis

Benefits of psychotherapy

Card one column	Value	Information
1–5	....	Study identification number
6	....	Running comparison number
7–8	....	Running measure number
9	....	Running record number: punch 1 for card 1
		Author
10–11	....	Publication date
12	....	Publication form: (1) journal, (2) book, (3) thesis, (4) unpublished
13	....	Training of experimenter: (1) psychology, (2) education, (3) psychiatry, (4) social work, (5) other, (6) unknown
14	....	Blinding: (1) E did therapy, (2) E knew composition of groups but didn't do therapy, (3) single-blind, (4) unknown
15	....	Did E call this an analogue study: (1) yes, (2) no
		Clients
16–17	....	Major diagnosis: (1) neurotic or complex phobic, (2) simple phobic, (3) psychotic, (4) normal, (5) character disorder, (6) delinquent or felon, (7) habituee, (8) mixed, (9) unknown, (10) emotional/somatic complaint, (11) handicapped, (12) depressive label
18–20	....	List code for label
21	....	Type of phobia: (1) reptile, (2) rodent, (3) insect, (4) speech, (5) tests, (6) other performance, (7) heights, (8) other
22–23	....	Average length of hospitalization in years
24	....	Average intelligence: (1) below average, (2) average, 95–105, (3) above average
25	....	Source of IQ: (1) stated, (2) directly inferred, (3) estimated
26	....	Similarity of client to therapist: (1) very dissimilar, (2) moderately dissimilar, (3) moderately similar, (4) very similar
27–28	....	Mean age to nearest year
29–30	....	Percentage male
31	....	SES: (1) low, (2) middle, (3) high, (blank) unknown
32	....	Solicitation of clients: (1) autonomous presentation, (2) presentation in response to advertisement, (3) solicited by E, (4) committed, (5) referred

Card one Column	Value	Information
		Design
33	....	Group assignment of clients: (1) random, (2) matching, (3) pretest equation, (4) convenience sample, (5) other nonrandom
34	....	Group assignment of therapists: (1) random, (2) matching, (3) nonrandom, (4) single therapist, (5) not applicable
35	....	Internal validity: (1) low, (2) medium, (3) high
36	....	Number of threats to internal validity
37-38	....	Percentage mortality from treated groups
39-40	....	Percentage mortality from comparison group
41	....	Is more than one therapy compared simultaneously against control: (1) yes, (2) no
42	....	Number of comparisons in this study
43	....	Number of this comparison
44-45	....	Number of outcome measures within this comparison
46-47	....	Number of this outcome measure (the rest of the record deals with this outcome measure)
		Treatment
48-49	....	Type of treatment: (2) placebo, (3) psychodynamic, (4) client-centered, (5) Adlerian, (6) gestalt, (7) systematic desensitization, (8) cognitive/Ellis, (9) cognitive/other, (10) transactional analysis, (11) behavior modification, (12) eclectic/dynamic, (13) eclectic behavioral, (14) reality therapy, (15) vocational/personal development counseling, (16) cognitive behavioral, (18) implosion, (19) hypnotherapy, (20) other
		Label for therapy type
		Proponent
50-52	....	List code for label
53-55	....	List code for proponent
56	....	Confidence of classification: (1) low . . . (5) high
57	....	Class of therapy
58	....	Superclass of therapy
59	....	Type of comparison: (1) control, (2) placebo, (3) second treatment
60	....	Type of control group: (1) no treatment, (2) waiting list, (3) intact group, (4) hospital maintenance, (5) other, (blank) not control
61-62	....	Type of placebo list code
		Label of placebo type
63-65	....	Second treatment type
66	....	Allegiance of E to therapy compared: (1) yes, (2) no, (3) unknown
67	....	Modality: (1) individual, (2) group, (3) family, (4) mixed, (5) automated, (6) other, (7) unknown
68-69	....	Location of treatment: (1) school, (2) hospital, (3) mental health center, (4) other clinic, (5) other outpatient, (6) private, (7) other, (8) unknown, (9) college mental health facility, (10) prison, (11) residential facility
70-72	....	Duration of therapy in hours
73-75	....	Duration of treatment in weeks
76-77	....	Number of therapists
78-79	....	Experience of therapists in years

Card two column	Value	Information
1–5	....	Study ID
6	....	Running comparison number
7–8	....	Running measure number
9	....	Running record number: punch 2 for card 2
		Effect size
10–12	....	Sample size for treatment group
13–15	....	Sample size for comparison group
16–17	....	Outcome type: (1) fear/anxiety, (2) self-esteem, (3) test measures and ratings of global adjustment, (4) life indicators of adjustment, (5) personality traits, (6) emotional/somatic disorders, (7) addiction, (8) sociopathic behaviors, (9) social behaviors, (10) work-school achievement, (11) vocational/personal development, (12) physiological measures of stress, (13) other
		Label of outcome measure
18–20	....	List code for outcome measure
21–23	....	Number of weeks post-therapy measure was taken
24	....	Reactivity of measure: (1) low . . . (5) high
25–26	....	Calculation of effect size: (1) mean difference over control S.D., (2) MS within, (3) MS total minus treatment, (4) probit, (5) chi square, (6) T table, (7) mean and P, (8) nonparametrics, (9) correlations, (10) raw data, (11) estimates, (12) other
27	....	Source of means: (1) unadjusted post-test, (2) covariance adjusted, (3) residual gains, (4) pre-post differences, (5) other
28	....	Significance of treatment effect: (0) −.001, (1) −.005, (2) −.01, (3) −.05, (4) −.10, (5) .10, (6) .05, (7) .01, (8) .005, (9) .001, (blank) not significant
29–34	....	Treatment group pre-mean
35–40	....	Treatment pre-standard deviation
41–46	....	Treatment post-mean
47–52	....	Treatment post-standard deviation
53–58	....	Comparison group pre-mean
59–64	....	Comparison pre-standard deviation
65–70	....	Comparison post-mean
71–76	....	Comparison post-standard deviation

Card three column	Value	Information
1–5	....	Study ID
6	....	Running comparison number
7–8	....	Running measure number
9	....	Running record number: punch 3 for card 3
10–13	....	T statistic
14–17	....	F statistic
18–22	....	Mean square within, residual, or common
23–24	....	Treatment group percentage improved
25–26	....	Comparison group percentage improved

Column	Value	Information
27–30	....	Effect size
31	....	Class of second therapy
32	....	Superclass of second therapy
33	....	Allegiance of E to second therapy
34	....	Modality of second therapy
35	....	Location of second therapy
36–38	....	Duration of second therapy in hours
39–41	....	Duration of second therapy in weeks
42–43	....	Number of therapists in second therapy
44–45	....	Experience of therapists in second therapy
46	....	Other factorial effects tested: (0) none,(1) race, (2) SES, (3)IE, (4) sex, (5) other
47	....	Is this the last effect with this comparison: (1) yes, (2) no
48–51	....	If yes, average effect size within this comparison
52	....	Is this the last effect size in this study: (1) yes, (2) no
53–56	....	If yes, average of all effect sizes in the study

APPENDIX B

STUDY USED AS CODING EXAMPLE IN CHAPTER FOUR

Appendix removed due to copyright restrictions. Material removed can be obtained as:

> Krumboltz, John D.; Thoresen, Carl E. The Effect of Behavioral Counseling in Group and Individual Settings on Information-Seeking Behavior. Journal of Counseling Psychology, v11 n4 p324-33, 1964.